Integrating neurobiological markers of depression: an fMRI-based pattern classification approach

Integration neurobiologischer Marker depressiver Erkrankungen mittels fMRT-basierter Musterklassifikation

Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences,

Julius-Maximilians-Universität Würzburg,

Section *Neuroscience*

submitted by

Tim Hahn

from Lennestadt

Würzburg

2010

Submitted on: April 1, 2010

Members of the *Promotionskomitee*:

Chairperson:	Prof. Dr. Dr. Martin J. Müller
Primary Supervisor:	Prof. Dr. Andreas J. Fallgatter
Supervisor (Second):	Prof. Dr. Klaus-Peter Lesch
Supervisor (Third):	Prof. Dr. Martin A. Heisenberg

Date of Public Defence: July 26, 2010

Date of Receipt of Certificates:

“Prediction is very difficult, especially about the future.”

Niels Bohr (1885-1962)

Table of Contents

0. Abstract

English

While depressive disorders are, to date, diagnosed based on behavioral symptoms and course of illness, the interest in neurobiological markers of psychiatric disorders has grown substantially in recent years. However, current classification approaches are mainly based on data from a single biomarker, making it difficult to predict diseases such as depression which are characterized by a complex pattern of symptoms. Accordingly, none of the previously investigated single biomarkers has shown sufficient predictive power for practical application.

In this work, we therefore propose an algorithm which integrates neuroimaging data associated with multiple, symptom-related neural processes relevant in depression to improve classification accuracy. First, we identified the core-symptoms of depression from standard classification systems. Then, we designed and conducted three experimental paradigms probing psychological processes known to be related to these symptoms using functional Magnetic Resonance Imaging. In order to integrate the resulting 12 high-dimensional biomarkers, we developed a multi-source pattern recognition algorithm based on a combination of Gaussian Process Classifiers and decision trees.

Applying this approach to a group of 30 healthy controls and 30 depressive in-patients who were on a variety of medications and displayed varying degrees of symptom-severity allowed for high-accuracy single-subject classification. Specifically, integrating biomarkers yielded an accuracy of 83% while the best of the 12 single biomarkers alone classified a significantly lower number of subjects (72%) correctly.

Thus, integrated biomarker-based classification of a heterogeneous, real-life sample resulted in accuracy comparable to the highest ever achieved in previous

single biomarker research. Furthermore, investigation of the final prediction model revealed that neural activation during the processing of neutral facial expressions, large rewards, and safety cues is most relevant for over-all classification. We conclude that combining brain activation related to the core-symptoms of depression using the multi-source pattern classification approach developed in this work substantially increases classification accuracy while providing a sparse relational biomarker-model for future prediction.

Deutsch

Während depressive Erkrankungen bislang größtenteils auf der Basis von Symptomen auf der Verhaltensebene und den jeweiligen Krankheitsverläufen diagnostiziert werden, hat das Interesse an der Verwendung neurobiologischer Marker bei psychischen Erkrankungen in den letzten Jahren stark zugenommen. Da jedoch die momentan verfügbaren Klassifikationsansätze zumeist auf Informationen eines einzelnen Biomarkers beruhen, ist die Vorhersage von auf der Symptomebene so komplexen Erkrankungen wie Depressionen in der Praxis deutlich erschwert. Dementsprechend konnte keiner der einzelnen bisher untersuchten Biomarker eine Vorhersagegüte erreichen, die für die praktische Anwendung eines solchen Ansatzes im klinischen Alltag ausreichend wäre.

Vor diesem Hintergrund schlagen wir deshalb zur Verbesserung der Klassifikationsgüte einen Algorithmus vor, der Messdaten vielfältiger depressionsrelevanter neuronaler Prozesse integriert. Zunächst wurden hierzu die Kernsymptome depressiver Erkrankungen aus standardisierten Klassifikationssystemen ermittelt. Anschließend entwickelten wir drei experimentelle Paradigmen, welche die Messung neuronaler Korrelate der mit den depressiven Kernsymptomen assoziierten psychologischen Prozesse mittels funktioneller Kernspintomographie ermöglichen. Um die resultierenden 12 hochdimensionalen Biomarker zu integrieren, entwickelten wir basierend auf der Kombination von Gauß-Prozess Klassifikatoren und Entscheidungsbäumen einen zweistufigen Mustererkennungsalgorithmus für multiple, hochdimensionale Datenquellen.

Dieser Ansatz wurde an einer Gruppe von 30 gesunden Probanden und 30 unterschiedlich schwer betroffenen und unterschiedlich medizierten stationären depressiven Patienten evaluiert. Insgesamt ermöglicht der Ansatz eine hohe Klassifikationsgüte auf Einzelfallebene. Insbesondere die Integration der

verschiedenen Biomarker führte zu einer Klassifikationsgüte von 83%, wohingegen die alleinige Klassifikationsgüte der 12 einzelnen Biomarker mit bestenfalls 72% deutlich geringer ausfiel.

Somit konnte der entwickelte Klassifikationsansatz in einer heterogenen, im Alltag aber typisch anzutreffenden depressiven Patientenstichprobe, eine Klassifikationsgüte erreichen, die mit der bislang bestmöglichen durch einzelne Biomarker erreichten Klassifikationsgüte in selektiven Einzelstichproben vergleichbar ist. Darüber hinaus zeigte die Analyse des empirischen Prädiktionsmodells, dass die Kombination der neuronalen Aktivität während der Verarbeitung von neutralen Gesichtern, großen monetären Belohnungen und Sicherheitssignalen zur optimalen Gesamtklassifikation führt. Zusammenfassend lässt sich schlussfolgern, dass der im Rahmen dieser Arbeit entwickelte, zweistufige Mustererkennungsalgorithmus für multiple, hochdimensionale Datenquellen die Klassifikationsgüte substantiell verbessert und erstmals die Konstruktion eines effizienten relationalen Biomarker-Modells für zukünftige Vorhersagen ermöglicht.

1. Introduction

Psychiatric disorders are currently diagnosed based on behavioral symptoms and course of illness according to standard classification systems such as DSM-IV (APA, 1994) or ICD-10 (WHO, 1992). Thus, specific behavioral and cognitive patterns – rather than etiology or pathophysiological mechanisms – are central to diagnosis. Over the last decades, the emerging consensus about the disorders and their symptoms has enabled scientific investigation of the respective disorders fostering the development of standardized instruments for diagnosis and the optimization of disease-specific treatments and evaluation protocols. In parallel, the understanding of the psychological processes associated with certain symptoms increased and technological advances made the investigation of the physiological underpinnings of such processes feasible (see 3.1.2 Biological markers of depression). In particular, genetic analyses and neuroimaging methods such as electroencephalography (EEG), positron emission tomography (PET), or functional magnetic resonance imaging (fMRI) have contributed greatly to our understanding of the biology of mental processes in humans, both normal and pathological.

Based on these developments, the interest in biomarkers of mental diseases has increased dramatically in recent years. In 2008 the number of relevant publications was more than eighteen times higher than in 2000 (Singh & Rose, 2009). The Biomarker Definitions Working Group (2001) has defined a biomarker as "*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention*". Thus, biomarkers might improve diagnosis of a particular medical condition or predict a patient's response to treatment enabling custom-tailored interventions. In addition, identifying biomarkers and investigating their

interrelations might, in the long run, help to uncover causal pathophysiological processes which may lead to physiology-based rather than behavioral-symptom guided nosological entities and enable custom-tailored therapeutic interventions.

Generally, biomarkers would be most needed in heavily disabling disorders for which diagnosis or prediction of treatment response is difficult. For depression with its high prevalence and high degree of suffering (see 3.1.1.1 Epidemiology of depression), biomarkers thus appear most valuable. While first suggestions to integrate biomarkers into standard diagnostic systems have been made (e.g. for bipolar disorder in DSM-V; Phillips & Vieta, 2007) and significant predictors of treatment response have been identified (Costafreda, Chu, Ashburner, & Fu, 2009; Costafreda, Khanna, Mourao-Miranda, & Fu, 2009), so far none have displayed sufficient predictive power for practical application (Singh & Rose, 2009; 3.2 Summary and goals of the study).

Against this background, we will develop a procedure which draws on clinical expertise, progress in the identification of correlates of mental disorders, and recent advancements in the field of pattern recognition alike to improve the predictive power of biomarkers of depression: Following this procedure, first, the core-symptoms of depression are identified from standard classification systems (3.1.1.2 Symptoms and diagnosis of depression). Then, three experimental paradigms probing psychological processes known to be related to these symptoms (3.1.2 Biological markers of depression) are designed and conducted acquiring task-related neuroimaging data from 30 depressive patients and 30 healthy control subjects (3.3 Materials and Methods). Finally, data from the resulting 12 biomarkers (3.3.3 Functional Magnetic Resonance Imaging) are integrated to allow for high-accuracy single-subject classification. As currently no method for the integration of the results of multiple high-dimensional biomarkers is available, we propose a principled

algorithm based on pattern recognition which is able to integrate information from multiple high-dimensional sources (Part I – Integrating biomarkers: development of a multi-source pattern classification algorithm).[1] Additionally, we will quantify the utility of each biomarker and investigate the model of the interrelations of those markers as well as the network of brain regions underlying prediction. Finally, we will discuss the resulting integrated biomarker model in the context of depression as well as directions of future research.

[1] Note that for ease of reading, we will describe the multi-source pattern classification algorithm developed in this work before outlining and discussing the study and its results.

2. Part I – Integrating biomarkers: development of a multi-source pattern classification algorithm

As mentioned above, no biomarkers have found their way into practical application, yet. To understand why psychiatry, in spite of tremendous progress over the last decades, has not succeeded in identifying biomarkers of sufficient predictive power, one needs to appreciate the vast complexity and heterogeneity of the physiological processes underlying psychiatric conditions: Virtually all mental disorders are assumed to be caused by complex interactions of personal experience, learning history, and individual personality traits as well as genetic and epigenetic factors. Thus, a large number of highly dynamic and interacting mechanisms – mainly within but not limited to the central nervous system – have to converge on various timescales to constitute a disorder. Hence, common techniques to measure potentially relevant physiological properties and processes suitable for use as biomarkers often yield very high-dimensional datasets. Typical results of genetic or neuroimaging analyses, for instance, can easily generate hundreds of thousands of individual measurements per person. In order to obtain a useful biomarker, a rule for prediction needs to be derived from datasets consisting of such measurements from a larger number of subjects.

Circumventing the problem of dimensionality, a vast body of research has identified potential biomarkers by considering subsets or statistical summaries of group data. Selection of such subsets or summary parameters is commonly guided by theoretical considerations specific to a disorder. As the aim of biomarker research is to identify characteristics relevant to individual predictions, studies have moved away from group data towards the evaluation of predictive power for single subjects (Caruana, Karampatziakis, & Yessenalina, 2008). In contrast to group analyses,

single-subject approaches predict characteristics of individuals (e.g. diagnosis or treatment response) by means of a previously learned rule rather than comparing (group) statistical parameters such as the mean of a group of subjects. Simple examples are psychometric tests for which a previously determined cut-off value enables classification of individual subjects (e.g. for depression: Beck, Steer, Ball, & Ranieri, 1996).

In recent years, methods which enable single-subject predictions not from a single value, but based on the aforementioned high-dimensional datasets have been developed and are increasingly used (Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005; Marquand, Mourao-Miranda, Brammer, Cleare, & Fu, 2008; Marquand, et al., 2009; Davatzikos, et al., 2005; Davatzikos, et al., 2005; Bode & Haynes, 2009; Haynes, 2009). These so-called pattern recognition algorithms have brought single-subject predictions based on genetic or neuroimaging data within reach. As one of the first, Davatzikos et al. (2005) successfully classified patients suffering from schizophrenia and healthy controls using a support-vector machine (SVM) on structural MRI data. Using a similar approach on genetic data, Huang and Kecman (2005) predicted colon and lymphoma cancer with significant accuracy.

Despite these advances, current pattern recognition approaches are mostly based on single biological markers, such as for instance genetic data or the neural responses related to a single pathologically deviating process alone. While first attempts to combine two sources of potentially clinically relevant data have been successful in the field of neuroimaging (Calhoun, Maciejewski, Pearlson, & Kiehl, 2008; Michael, Calhoun, Andreasen, & Baum, 2008) and for the prediction of breast cancer based on genetic data (Zeng & Liu, 2009), to date, a principled method integrating results from multiple classifiers is not available. Considering the fact that all psychiatric disorders are diagnosed based on multiple symptoms associated with

a potentially large number of physiological processes, this appears conceptually unsatisfying and methodologically suboptimal. In the following, we will therefore propose a principled procedure integrating information from multiple high-dimensional biomarkers in order to allow for a more comprehensive, biomarker-based classification of psychiatric disorders. While we developed and implemented the algorithm for use with neuroimaging data and provide examples from this context, it can be used to identify and be applied to biomarkers of any data type and scale including mixed analyses of genetic, psychometric, and neuroimaging data.

2.1. Approaches to classification

Basically, classification is analogous to regression when the variable predicted is discrete. From this point of view, classification is the prediction of class labels from the data. It follows that a classifier is a function that predicts the class label of the class to which a sample (i.e. for instance a subject) belongs based on the values of the features of that sample ($\mathbf{x}_i$; i.e. for instance the voxel values). In order to be able to predict a class label, a function f mapping the data to the labels needs to be found. In other words, we need to model the relationship between a sample's features and its class label. The resulting model f which predicts a class label from a sample is called a classifier. From a geometric point of view, one can consider the samples as points in a space where each feature corresponds to one dimension (feature space) – e.g. a three dimensional space, if three measurements have been made. Then, classification corresponds to finding a decision boundary – e.g. a plane in the case of

three features – which separates the classes[2]. Such a decision boundary is equivalent to the function f which maps the data to the labels.

Traditional approaches

When attempting to integrate multiple biomarkers to obtain a prediction – that is find a classifier f which maps the biomarker data to the labels (patients, controls) – the greatest challenge lies in the enormous number of features per sample (i.e. the amount of data per person). A single fMRI whole-brain volume can easily consist of 150,000 measurements (voxels). Traditional methods such as for example discriminant analysis (and all methods derived from it, such as multiple regression analysis), cannot handle this number of features: Apart from a multitude of other requirements which may or may not be met by the dataset, discriminant analysis cannot be used if the features (e.g. the voxel values) are redundant. As this would obviously be the case for neighboring voxels, the thus present multicollinearity leads to an ill-conditioned predictor matrix which cannot be inverted. Thus, no estimation would be possible. While the common solution to reduce dimensionality by selecting specific (potentially linearly independent) features has been used successfully in this context and in pattern recognition in general, this approach requires hypotheses about the data (e.g. regions of interest in the brain) which might not always hold true or be inappropriate for the question at hand. This is particularly problematic in the context of a new methodological development such as the one presented in this work (for an overview of methods for feature selection or dimensionality reduction, see Pereira, Mitchell, & Botvinick, 2009). In summary, traditional methods such as those

[2] Note that the decision boundary does not need to be a plane in feature space, but can take infinitely many shapes which correspond to a plane in hyperspace. For the sake of simplicity and visualization, we assumed a linear kernel function which would restrict the decision boundary to the shape of a plane in feature space. In those cases, the mapping function f is also linear. In this work, linear kernels/ covariance functions are used as they best avoid overfitting while enabling a valid mapping procedure (see 2.3.4 Multivariate feature mapping).

from the family of discriminant analyses can either not be used for the classification of extremely high-dimensional data or require substantial, often hypothesis-driven and thus subjective data preparation steps.

Pattern recognition approaches

More recent developments in the field of pattern recognition can, however, classify datasets of higher dimensionality and do not put constrains on the properties of the data: Generally, pattern recognition is a field within the area of machine learning which is concerned with automatic discovery of regularities in data through the use of computer algorithms. Using these regularities, it can classify data into different categories (Bishop, 2007). For instance, in the context of neuroimaging, brain images are treated as spatial patterns, and pattern recognition approaches are used to identify statistical properties of the data based on which the two groups of subjects (e.g. patients and controls) can be discriminated.

A classifier based on pattern recognition is first trained by providing examples of the form $< \mathbf{x}, c >$ where $\mathbf{x}$ represents a spatial pattern (i.e. the features) and c is the class label (e.g. $c = -1$ for patients and $c = +1$ for controls). Each spatial pattern (e.g. whole brain image) corresponds to a point in feature space. During the training phase, the pattern recognition algorithm estimates the mapping function (f) corresponding to the hyperplane which optimally separates the samples in feature space according to the class label. Once the decision function is determined from the training data, it can be used to predict the class label of a new, previously unseen sample. In this context, it is essential to obtain a decision function that not only classifies the training data correctly, but also does the same for the test data. One needs to be aware that a classifier which perfectly predicts the labels of the training data is by no means guaranteed to predict the labels of a new dataset correctly. The

term “overfitting” is used to refer to cases where the model fits the training data very well, but performs badly on new data. Thus, the accuracy of a classifier always needs to be evaluated using data which was not seen by the classifier before (for details, see *Probability prediction and accuracy estimation* at the end of 2.3.1 First-level prediction and *Class membership prediction and accuracy estimation* at the end of 2.3.2 Second-level prediction).

The most widely known method of pattern recognition is the support-vector machine (SVM; Vapnik, 1995) algorithm. As described above, binary classification can be understood in terms of finding a decision boundary which optimally separates the two classes. Directly using this idea, SVMs work by constructing the maximum margin hyperplane which finds the single hyperplane with the maximum distance between the plane and the points closest to the plane (the so-called support vectors). This simple approach guarantees an optimal separation of the dataset given a defined kernel function. However, the predictions of SVM are categorical. Hence, they do not provide probabilities or confidences associated with the classification scores. Basically, the output of an SVM classifier consists of a vector containing binary class labels (e.g. -1 for class A and 1 for class B).

Gaussian Process (GP) classifiers (Rasmussen & Williams, 2006) are another example of a pattern recognition classifier. While performing with comparable accuracy on neuroimaging data (Marquand, et al., 2009), they provide probabilistic class label predictions. GP classifiers are based on Bayesian probability theory and are therefore guaranteed to handle probability distributions correctly. GP classifiers are most easily understood as a distribution over functions. GP inference consists of applying Bayes’ rule to find the (posterior) function distribution that best approximates the training data. Specifically, GP classification is an extension of the GP regression model in which data are classified by applying a latent regression model, which is

then constrained to the unit interval to produce probabilistic predictions (for details, see 2.3.1 First-level prediction).

2.2. Goals and challenges of algorithm development

Goals of algorithm development

The main goal of algorithm development is the construction of a procedure which allows for single-subject classification based on multiple high-dimensional biomarkers such as those commonly obtained using neuroimaging or genetic analyses. Integrating a potentially large number of biomarkers, it is also of interest – for research as well as for potential application – which biomarkers contributed in what way to classification (e.g. which genes or brain regions play which role for prediction). From this, we derive four specific aims:

1) The algorithm must be able to classify groups based on multiple, high-dimensional datasets (i.e. biomarkers).
2) In order to efficiently identify those datasets holding maximum predictive power, it must be possible to quantify the contribution of each biomarker to over-all prediction.
3) As is already possible today, we want to quantify the contribution of each biomarker's single features to classification. This enables the identification of the most discriminative properties of the single dataset.
4) Combining multiple biomarkers, we want to quantify the contribution of single features to over-all classification. In contrast to (3), this would enable the identification of the most discriminative properties of a biomarker in the context of all other biomarkers.

As it greatly enhances practical applicability, we will use only methods which could be applied to biomarkers of any level of measurement and data type, including neuroimaging, genetic, and psychometric data.

Challenges of algorithm development

As outlined above, methods of pattern recognition developed in recent years are able to classify high-dimensional datasets with high accuracy. A growing body of evidence mainly from neuroimaging underlines this (Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005; Marquand, et al., 2008; Marquand, et al., 2009; Davatzikos, et al., 2005; Davatzikos, et al., 2005; Bode & Haynes, 2009; Haynes, 2009). New challenges arise, however, when aiming to integrate data not from one, but from multiple biomarkers: For prediction, at least one dataset (e.g. a whole-brain volume) per biomarker has to enter the analysis for each subject. If, for instance, only five whole-brain volume scans which might be potential biomarkers are considered, this amounts to at least 750,000 measurements per person which corresponds to a feature space with 750,000 dimensions. While methods of pattern recognition have yielded very good results in recent years, mostly single brain volumes have been considered so that no data on classification performance is available for ultra-high dimensional datasets as they emerge when combining biomarkers. From a theoretical perspective, increasing the number of dimensions of feature space beyond the already challenging numbers currently used, might be problematic (for an empirical evaluation of high-dimensional methods, see Caruana, et al., 2008).

Even disregarding these mathematical issues which might in the long-run be solved using more powerful kernel combination methods or automatic feature selection approaches, the problem of interpretability arises when combining datasets (i.e. biomarkers): While at first sight, it might seem desirable to base prediction on all

features irrespective of their source, this procedure would identify a pattern of discriminative features across biomarkers. This entails that the contribution of one biomarker to over-all classification would only have meaning in the context of the other biomarkers which were entered in the analysis. In the field, classification would then rely on a large number of highly complex, non-linear, multivariate interactions of for example brain regions measured during task A and (other) regions measured during task B and C. Furthermore, removing even a single biomarker – i.e. a subspace of the over-all feature space – would change the discriminative pattern learned by the classifier and alter predictions for new data in a completely unforeseeable manner. This would make the independent evaluation of single biomarkers within the set impossible. Furthermore, it would not allow for a quantification of the contributions of each biomarker's single features to classification.

2.3. Algorithm development

Addressing these issues, we provide a solution based on a two-level procedure: On the first level, we use pattern recognition classifiers on each biomarker independently (2.3.1 First-level prediction). For b biomarkers, this creates b classifiers acting independently in each of the b feature spaces. Thereby, the number of dimensions to be considered by each classifier is limited to the maximum number of features for a single biomarker (d). Thus, the available pattern recognition algorithms ought to be able to provide reliable and accurate single-subject predictions for each single biomarker independently. Analysis on the second level is then based on the matrix containing the predictions made by each classifier for each person. This way, the ultra-high dimensional classification problem is reduced to a lower-dimensional space. For b datasets containing d features, the problem is

reduced from a $b \times d$-dimensional space to a b-dimensional space. For $m = 20$ subjects and $b = 10$ biomarkers containing $d = 150{,}000$ features each, for instance, the problem is reduced from 20 objects in a 1,500,000-dimensional space to 20 objects in a 10-dimensional space. The challenge for second-level classification now lies in the fact that the new feature space is by orders of magnitude more densely occupied than the original space, rendering linear decision functions inappropriate while non-linear classifiers run the risk of overfitting while simultaneously hampering a straightforward interpretation of the results. With a decision tree algorithm, a powerful while easily interpretable classifier is chosen to address these issues (for the choice of method and details on tree classification, see 2.3.2 Second-level prediction).

2.3.1. First-level prediction

As outlined above, pattern recognition algorithms can be used to obtain single-subject predictions from high-dimensional datasets. GP classifiers appear particularly well suited for the two-level approach (i.e. for producing input for second-level classification outlined in section 2.3.2 Second-level prediction) since GP output consists of class probabilities which – in contrast to the binary outputs of other pattern recognition classifiers such as SVM – preserve a maximum of information.[3] As the performance of GP classifiers is comparable to that of SVM (Marquand, et al., 2009) while mapping feature weights (for details, see 2.3.4 Multivariate feature mapping) is more flexible, we will use GP classifiers for all first-level predictions.

[3] Note that procedures have been suggested to transform SVM output into class probabilities. This does, however, require a nested-cross-validation procedure which is prone to overfitting and can thus be unstable in smaller samples and nested subsamples. Furthermore, these approaches rely basically on Bayesian posterior probabilities – a feature already inherent in all GP methods (Dai, Srikant, & Zhang, 2004).

In the following, we will outline the basic mathematical ideas necessary to understand how first-level predictive probabilities for second-level classification are obtained (for an in-depth introduction to Gaussian Processes including fundamental proofs related to equations outlined in this section, see Rasmussen & Williams, 2006). Basically, a GP classifier can be seen as a function which predicts a subject's probability to be a member of class -1 based on a multivariate pattern within that subject's features. As GP classification is an extension of the GP regression model, we will first describe GP regression to then show how GP classification probabilities can be obtained from this. Finally, we outline how learning – i.e. the estimation of certain GP parameters from a training dataset – is conceptualized.

Gaussian Process Regression

As in all regression and classification problems, we begin with a set of training data $D = \{\mathbf{x}, y\}$ where $\mathbf{x}$ is an $m \times d$ matrix (*m* training samples with *d* features each) consisting of input vectors $\mathbf{x}_i$ while y is a column vector of target variables where $y_i \in \{+1,-1\}$ for binary classification (for multi-class classification, see Rasmussen & Williams, 2006; for regression $y_i \in \Re$). Training samples are indexed by $i = 1,...,m$. As outlined above, a classifier can be seen as a function of the features which allows for an accurate prediction of a target $y*$ from a previously unseen sample $\mathbf{x}*$ (i.e. not the training data). For Gaussian Process Regression (GPR), this is an estimate of the target variable while for binary Gaussian Process Classification (GPC) predictions consist of class probabilities. In this work, class probabilities for each subject are calculated as the conditional probability to be in class -1 (which will later denote the patient group; see 3.3.1 Participants) given the previously seen samples and the new

sample: $p(y^* = -1 \mid \mathbf{x}^*, D)$. For GPR as well as for GPC, inferences are made in accordance with Bayesian probability theory (see below).

Mathematically, a Gaussian process (GP) is defined as the generalization of the multivariate Gaussian distribution to infinitely many dimensions with the constraint that drawing examples from any finite dimensional subspace must always yield a multivariate Gaussian distribution. In much the same way that drawing examples from the well known univariate Gaussian distribution will always yield a Gaussian distribution, drawing from a GP will always yield a multivariate Gaussian distribution. While a Gaussian distribution can be defined by its mean and variance (i.e. its mean vector and covariance matrix for multivariate Gaussian distributions) a GP is uniquely described by its mean and covariance functions ($GP \sim N(m(\mathbf{x}), k(\mathbf{x}_i, \mathbf{x}_j))$).

Against this background, GPR can basically be seen as a Bayesian extension of a simple regression model $y = \mathbf{x}^T \mathbf{w} + \varepsilon$, where $\mathbf{w}$ is a vector of weights and $\varepsilon \sim N(0, \sigma_n^2)$ is a Gaussian noise term. In Bayesian statistics, so-called posterior probabilities (probabilities taking into account knowledge we have gained considering the training samples) are basically calculated from prior probabilities (only considering knowledge we had before, i.e. none when calculation starts which corresponds to a zero-mean prior) and their respective conditional probabilities using Bayes' theorem.[4] Colloquially, it could be stated as

$$p(prediction \mid data) = \frac{likelihood * prior}{normalization\_term},$$

where $p(prediction \mid data)$ is the probability of the prediction after we have taken the prior and the likelihood of the data into account (normalized by a constant). In a GP

[4] In case of ordinary least square regression, $\mathbf{w}$ would be estimated simply by $\mathbf{w} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T y$.

model, a zero-mean GP prior is placed over the weights before the posterior distribution is computed by

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \theta) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \theta) * p(\mathbf{w} \mid \theta)}{p(\mathbf{y} \mid \mathbf{X}, \theta)} \qquad \text{equation 1}$$

Here, the vector $\mathbf{y} = [y_1, ..., y_m]^T$ denotes the targets (labels for classification), $p(\mathbf{w} \mid \theta)$ describes the prior, the likelihood is denoted by $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \theta)$, and θ is a vector of hyper-parameters (for the estimation of the hyper-parameters, see *Learning in GP models* below). The denominator is called the marginal likelihood (or model evidence) and can be expressed as $p(\mathbf{y} \mid \mathbf{X}, \theta) = \int p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$. Thus, the posterior probability of a GP model is computed by calculating the product of the likelihood and the prior and then dividing it by the model evidence.

When making a prediction for a previously unseen sample, we integrate (average) over all possible values for **w**, weighted by their posterior probability: $p(f^* \mid D, \theta, \mathbf{x}^*) = \int p(f^* \mid \mathbf{w}, \mathbf{x}^*, \theta) p(\mathbf{w} \mid D, \theta) d\mathbf{w}$. Thus, GP predictions are a weighted average of all possible linear models[5] under the prior assumptions, based on the samples that have already been seen.

Alternatively, we can understand the construction of a GP classifier as the selection of a function $f(\mathbf{x})$ which maps the data to the labels (i.e. to the values for GPR). Viewing a GP as a distribution over functions (a process "containing infinitely many functions"), we can select the most likely function given the samples we have seen. To make a prediction for a new sample, we again place a zero-mean GP prior

[5] Analogues to our consideration regarding the shape of the SVM hyper-plane (2), the model in this case does not need to be linear. Its form depends on the choice of the kernel or the covariance function in GP models. Here, a linear covariance function is used as they best avoid overfitting while enabling a valid mapping procedure (see 2.3.4 Multivariate feature mapping).

over the distribution of functions, and then use Bayes' rule to determine the posterior distribution evaluated at the training data (Figure 1).

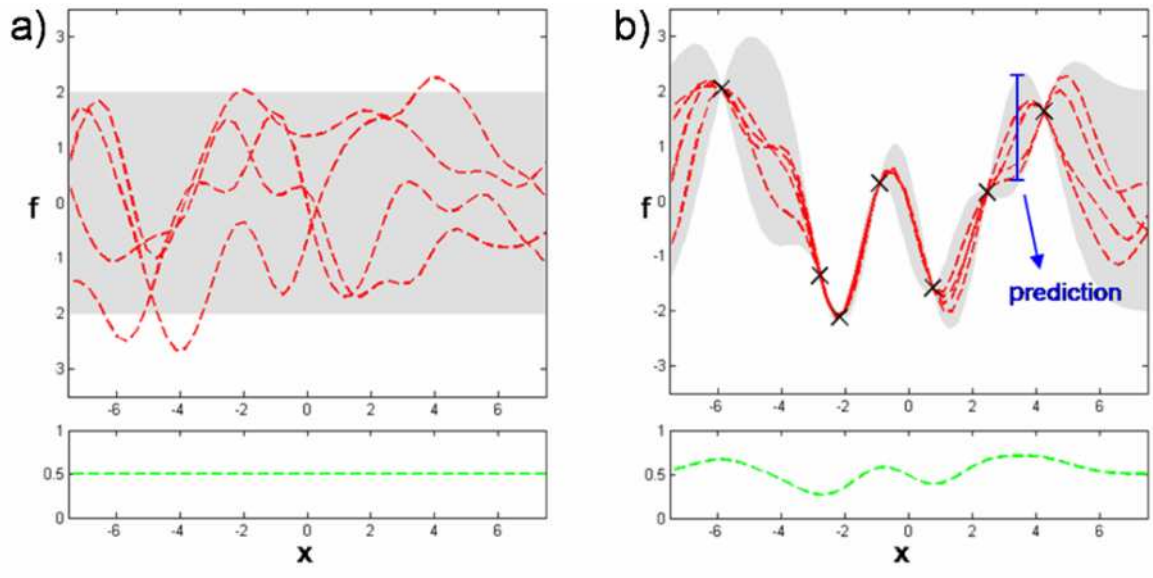

Figure 1. Schematic representation of Gaussian process prediction from one-dimensional data (x). (a) Sample functions drawn from the prior distribution. (b) Sample functions drawn from the posterior distribution after data have been added (black crosses). Top panels show values of the (latent) function and bottom panels show the function after squashing it through the probit likelihood [to obtain classification probabilities; see *Gaussian Process Classification*]. Grey shaded areas indicate 95% confidence intervals and one test data point is shown in blue (from Marquand, et al., 2009).

An attractive feature of GPR models is that the likelihood and the prior are both GPs. It follows that the posterior distribution is also Gaussian. Thus, the mean and variance uniquely defining it can be computed in closed form:

$$p(f^* \mid D, \theta, \mathbf{x}^*) \sim N(\mu, \sigma^2)$$

$$\mu = \mathbf{k}^{*T}\, \mathbf{C}^{-1}\mathbf{y} \qquad \text{equation 2}$$

$$\sigma^2 = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{*T}\, \mathbf{C}^{-1}\mathbf{k}^*$$

where $\mathbf{C} = \mathbf{K} + \sigma_n^2\mathbf{I}$. **K** is a kernel matrix describing the covariance between each data sample (i.e. $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j, \theta)$). Similarly, $\mathbf{k}^* = [k(\mathbf{x}_1, \mathbf{x}^*), ..., k(\mathbf{x}_m, \mathbf{x}^*)]^T$ is a vector of covariances between the test (**x***) and training data (**X**). There are a number of possible forms for the covariance function (Rasmussen & Williams, 2006) but here we use a linear covariance function which is parameterized as:

$$\mathbf{K} = \tfrac{1}{l^2}\mathbf{X}\mathbf{X}^T + o \qquad \text{equation 3}$$

l is a length-scale parameter that controls how rapidly predictive variance grows with increasing distance from the data points. As one would intuitively expect, predictive variance is large in intervals of values for which little data is available. o is a bias term accommodating the offset from zero similar to the constant in linear regression (for a description of how the hyper-parameters are calculated, see *Learning in GP models*).

Gaussian Process Classification

GPC is an extension of GPR and predicting class labels is done by placing a GP prior over an unconstrained latent function and computing its posterior distribution. Thereby, the GPR predictions are constrained to the unit interval [0 1]. While there are several possibilities for this latent function, in this work we use the probit likelihood $\Phi(x)$ as it ensures that an approximation of the posterior by a Gaussian is

appropriate (see below; for a detailed discussion of how probability outputs are obtained using a latent function, see Rasmussen & Williams, 2006, chapter 3).

Exact inference for GP classification is not analytically tractable (i.e., it cannot be computed in closed form as possible for the regression model) because after applying the latent function, the likelihood and the posterior are no longer Gaussian. Instead, the posterior needs to be approximated by a Gaussian. We use the expectation propagation (EP) algorithm which has been shown to have superior performance to other approximation methods (Kuss & Rasmussen, 2005; Nickisch & Rasmussen, 2008; for a detailed description of EP, see Rasmussen & Williams, 2006, pp. 52-60). Finally, predictions can be made by integrating over the approximated posterior distribution.

Learning in GP models

Learning in a GP model refers to finding the best functional form and hyper-parameters for the covariance function given the training data (which given the zero mean is sufficient to uniquely define a GP). This is commonly done by maximizing the logarithm of the marginal likelihood which, for GPR, can be computed in closed form:

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\tfrac{1}{2}\mathbf{y}^T \mathbf{C}_\theta^{-1}\mathbf{y} - \tfrac{1}{2}\ln\left|\mathbf{C}_\theta\right| - \tfrac{n}{2}\ln 2\pi \qquad \text{equation 4}$$

where $\mathbf{C}_\theta$ refers to the evaluation of $\mathbf{C}$ given hyper-parameters θ.[6]

The marginal likelihood corresponds to the total probability of the data given the model hyper-parameters. It has the attractive property that it constitutes a trade-off

[6] Again, in GPC models, exact computation of the marginal likelihood is analytically intractable (again because of the non-Gaussian likelihood). However, it can be approximated using EP. Computationally, we implemented this by minimizing $-(\ln p(\mathbf{y} \mid \mathbf{X}, \theta))$ which is equivalent.

between good fit to the data and a penalty for model complexity. Thus, simpler models are favored which provides protection against overfitting. In this work, we use the marginal likelihood to select hyper-parameter values as well as a linear covariance function for prediction. The main reason for this is that the use of a linear kernel permits the straightforward construction of a multivariate mapping representation (2.3.4 Multivariate feature mapping).

Probability prediction and accuracy estimation

In order to predict a subject's probability to be in class -1 (i.e., to be a patient in the later study) a leave-one-out cross-validation (LOO-CV) is done during which this subject is excluded from the GPC training data (i.e. this sample is not considered during the estimation of the hyper-parameters).[7] Thus, the procedure ensures that the classifier never knows the true label of the subject whose probability is to be predicted. This way, we avoid high accuracy rates simply due to overfitting. The LOO-CV yielding the vector containing each subject's probability to be in class -1 (p_{GP}) is conducted as follows: In each leave-one-out run, we use data from all but one subject per group (*S*-1 of the *S* subjects per group) to train the classifier. Subsequently, the probability to be in class -1 of the remaining pair of subjects (one patient and one control), which was so far unseen by the algorithm, is calculated. This procedure is repeated *S* times, each time leaving out a different pair of subjects, yielding each sample's probability to be in class -1 for each biomarker. Single classifier accuracy is calculated as the ratio of correct predictions over number of cases.

[7] For ease of reading, we will refer to the vector containing each subject's predicted probability to be in class -1 calculated in accordance with this procedure as p_{GP} for the rest of this work.

2.3.2. Second-level prediction

By using single GP classifiers on each of the biomarkers, we have reduced the number of dimensions of the over-all problem space to the number of biomarkers. Basically, this corresponds to a projection of each single biomarker's feature space onto one dimension. Combining these dimensions, a new, lower-dimensional problem space is created in which final classification must now be performed. Choosing a suitable second-level classification algorithm from the large number of possibilities (e.g. discriminant analyses, Naïve Bayesian Classifiers, kernel density estimations, k^{th} nearest neighbors, SVM, or another GPC, to name only a few) a number of fundamental issues need to be considered: Second-level space is very densely occupied in comparison to the first-level spaces. For example, for 20 subjects and 10 common neuroimaging biomarkers with 150,000 dimensions each, the ratio of objects to dimensions increases from $\frac{1}{75000}$ to 2. While this is generally desirable, a linear decision boundary might no longer suffice to separate the classes in this case. However, non-linear classifiers will not readily allow for an interpretation of the contribution of a single biomarker to final classification as the weights of the respective functions (or the projections on the weight vectors) are inherently multivariate, i.e. interdependent. In order to keep results transparent and interpretable – also regarding the later mapping procedure (2.3.4 Multivariate feature mapping) – while benefiting from the power of a non-linear classifier, we adopt a decision tree approach developed by Breiman et al. (1984).

The Classification and Regression Tree (CART) algorithm

Basically, the CART algorithm determines a set of if-then logical conditions that allow for the classification of subjects. Those conditions define for each relevant

variable (for how to determine this relevance, see below) a threshold value. If a subject's value on this variable is lower than the threshold, it is classified in one group, if it is equal to or higher than the threshold it is placed into the other.

Specifically, in the first step, the entire dataset (in our case the $m \times b$ matrix containing the p_{GP}-vector for each biomarker) is split into two subsets based on the single variable which produces the most homogeneous subsets in terms of class labels.[8] Two (more homogeneous) subsets of the original data are thus created. This corresponds to creating two subspaces of the original feature space. Each of the two resulting datasets is again split based on the variable which produces the most homogeneous subsets in terms of class labels for each respective subset. The process continues yielding four subsets of the original data, then eight and so on. Each new subset is more homogeneous in terms of class labels than the one from which it was created. In classification tree terminology, a "tree"-structure with two "branches" growing from each "node" is created. At each node, we find the variable which, split at the optimal value, creates two branches so that the resulting two nodes are maximally homogeneous or "pure". With this procedure, the feature space is partitioned into increasingly many subspaces which are increasingly pure in that they contain more and more subjects carrying the same label.

Following Breiman et al. (1984), we determine the impurity i of a node t based on Gini's diversity index. For binary classification, it can be calculated by

[8] Note that Breiman et al. (1984) have also suggested trees which use linear combinations of variables to split each node. For reasons of interpretability, we choose to use their initial suggestion and do not consider splits based on combinations of variables. For the same reason, we do not employ the related procedures of tree boosting which base classification upon combining trees obtained by repeatedly re-sampling subsets of the data. For testing purposes, we nonetheless used stochastic gradient boosting (Friedman, 2002) on the data obtained in "Part II – Classification in the context of depression". Due to enormous overfitting – as is often seen with increasingly powerful methods – accuracy rates did not improve.

$$i(t) = 1 - [p(y = -1 \mid t)^2 + p(y = 1 \mid t)^2] \quad \text{equation 5}$$

where $p(y = -1 \mid t)$ and $p(y = 1 \mid t)$ denote the proportion of subjects from group 1 and group 2, respectively, relative to the total number of subjects in t. Based on this, the variable selected for the next split must be the one which reduces $i(t)$ most. Obviously, this corresponds to maximizing $i(t_{parent}) - i(t_{children})$, where $i(t_{parent})$ is the impurity of the node to be split and $i(t_{children})$ is the impurity of the resulting two nodes.

The main problem to be solved by the algorithm is to determine when to stop growing the tree (i.e. splitting nodes further). If we set no stopping criteria, the tree will grow, reaching increasingly high classification accuracy until all samples in a node carry the same label (and are thus classified correctly) or until there is no meaningful information left which could decrease impurity of the nodes. While potentially performing extremely well on a given dataset, such a tree would be very likely to misclassify new samples. That is, it would be highly prone to overfitting. In order to avoid this, splitting nodes is stopped if Gini's splitting criterion based on $\max(i(t_{parent}) - i(t_{children}))$ is no longer fulfilled (for details on the implementation of the criterion, see Breiman, 1996) and/or the number of samples in t_{parent} is lower than 10.

In summary, we use GP classification probabilities (p_{GP}-vectors for each biomarker) as predictors based on which the algorithm constructs a classification tree. This corresponds to finding those variables most useful for the partitioning of the dataset into purer subsets and assigning a set of if-then logical conditions that allow for the classification of subjects in each subsample. This procedure is inherently non-linear – allowing for high flexibility in the densely occupied p_{GP}-vector space – while enabling simple interpretation and identification of the variables relevant for the classification within each subspace. While in the context of this work we construct the

tree from the p_{GP}-matrix containing class probabilities, it is easily possible to include variables of any level of measurement such as psychometric scores or low-dimensional genetic data.

Decision tree visualization and interpretation

To visualize the structure of the tree, we calculate a tree based on all subjects using the procedure described above. Note that samples with a value lower than the threshold for this variable are depicted as the left branch whereas samples with a value equal to or higher than the threshold for this variable are shown as the right branch.

We call the resulting structure the “optimal tree” as it considers the data from all subjects (in contrast to the tree generated using the leave-one-out procedure described below in *Class membership prediction and accuracy estimation*). Additionally, for a node t_{parent} to be split, we require that each of the nodes $t_{children}$ potentially generated contain no less than 10% of the samples. This is done to avoid displaying practically irrelevant nodes. Also it ensures a reasonable number of samples for the generation of node-specific multivariate maps (see 2.3.4 Multivariate feature mapping).

Class membership prediction and accuracy estimation

In order to calculate the overall prediction accuracy of this approach while avoiding high accuracy rates simply due to overfitting, a leave-one-out procedure is implemented in analogy to the one used to determine the accuracy of the single GP classifiers. Note that we use the same LOO-CV structure in the single classifiers and in the decision tree, which ensures that at each cross-validation fold, the test set is completely independent of the training set. In each leave-one-out run, we use the

predictive probabilities from all but one subject per group (*S*-1 of the *S* subjects) to train a decision tree model. Subsequently, the class memberships of the remaining pair of subjects are calculated based on the training tree model. This procedure is repeated *S* times, each time leaving out a different pair of subjects, yielding each sample's predicted overall class membership. Again, accuracy is calculated as the ratio of correct predictions over number of cases.

2.3.3. Significance Testing

To establish whether the observed single GP classification accuracies are statistically significant, we run each GP classifier 1000 times with randomly permuted labels and count the number of permutations which achieved higher accuracy than the one observed with the true labels. The p-value is then calculated by dividing this number by 1000.

In order to test whether the combination of data sources results in substantially increased classification accuracy compared to the accuracy obtained from the most informative of the sources alone, we proceed as follows: First, we obtain an estimate of the expected best single GP classification accuracy under permutation. This is done by running each GP classifier independently for all biomarkers with randomly permuted labels and taking the maximum accuracy. Doing this 1000 times provides a distribution of maximum accuracy under permutation. The median of this distribution constitutes the best estimate for the expected maximum single GP classification accuracy under permutation. Secondly, we re-run each GP classifier independently for all biomarkers with randomly permuted labels. This time, however, we calculate the accuracy of the decision tree based on the predictive probabilities derived with the randomly permuted labels. Doing this 1000 times provides a distribution of

decision tree accuracy under permutation. Subtracting the best estimate for the expected maximum single GP classification accuracy under permutation calculated above from this distribution creates the distribution of the expected difference between decision tree accuracy and single best accuracy under permutation. As the null hypothesis is that the decision tree does not substantially outperform the best individual classifier, the p-value is then calculated by counting the number of times that this expected difference under permutation exceeds the difference between the decision tree accuracy and the single best GP classification observed with the true labels and dividing it by 1000.

2.3.4. Multivariate feature mapping

In the following, we will outline how quantification of the contribution of each biomarker's single features to classification is commonly evaluated for single GP classifiers (see *Decision boundary weight and group distribution mapping*). This enables the identification of the most discriminative properties of a single dataset. With regard to the new multi-source classification algorithm developed in this work, we will then propose a way to identify the most discriminative properties of each biomarker in the context of over-all classification, i.e. within the optimal decision tree (see *Node-specific group distribution mapping*).

Decision boundary weight and group distribution mapping

For single GP classifiers, it is possible to obtain two representations of the contributions of the single features to classification. In both cases, this is achieved by calculating a linear combination of the subjects' feature vectors $\mathbf{x}_i$ (i.e. all data). The

methods differ mathematically only regarding the coefficients (weights) used to calculate the linear combination of the data.

The first method – decision boundary weight w-mapping – is analogue to the weight vector in SVM discriminant mapping. The so-called weight-vector $\mathbf{w}$ is orthogonal to the hyperplane separating the two classes.[9] For each feature, its value represents the contribution of this feature to the construction of the hyperplane (Mourao-Miranda, et al., 2005). It is calculated by

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i = \frac{1}{l^2} \mathbf{X}^T (\mathbf{C}^{-1}\mathbf{y}) \qquad \text{equation 6}$$

where α_i is the predictive mean (specifically, it consists of the last two weighting coefficients derived from GPC training with all samples; for details, see Marquand, et al., 2009). As one would intuitively expect considering the concept of decision boundary construction from the adjacent support-vectors in the SVM approach, samples closer to the hyperplane carry higher weights $\mathbf{w}$. Being simply a linear combination of the sample data weight by the predictive mean α_i, the $\mathbf{w}$-map basically quantifies a feature's contribution to the construction of the hyperplane (i.e. the decision boundary) considering all other features. Thus, it is a multivariate representation of the decision boundary.

Maps generated with the second method – group distribution g-mapping (Marquand, et al., 2009) – are calculated by weighting each sample by the mean of the (latent) function (μ; from equation 2) at each training point:

[9] Using a GP classifier, there obviously is no hyperplane in the regular sense. As noted above, however, any decision function of a binary classifier f can be described as a plane separating the two classes. To allow intuitive visualization, we adopt the feature space view in this section.

$$\mathbf{g} = \sum_{i=1}^{m} \mu_i \mathbf{x}_i = \mathbf{X}^T \mathbf{m} \qquad \text{equation 7}$$

Unlike $\mathbf{w}$, $\mathbf{g}$ quantifies the distribution of the classes with respect to each other. Thereby, it is a multivariate quantification of the difference between the predicted groups on a given dimension (feature). Figure 2 shows a schematic illustration and a geometric interpretation of $\mathbf{w}$ and $\mathbf{g}$ in a hypothetical two-dimensional feature space.

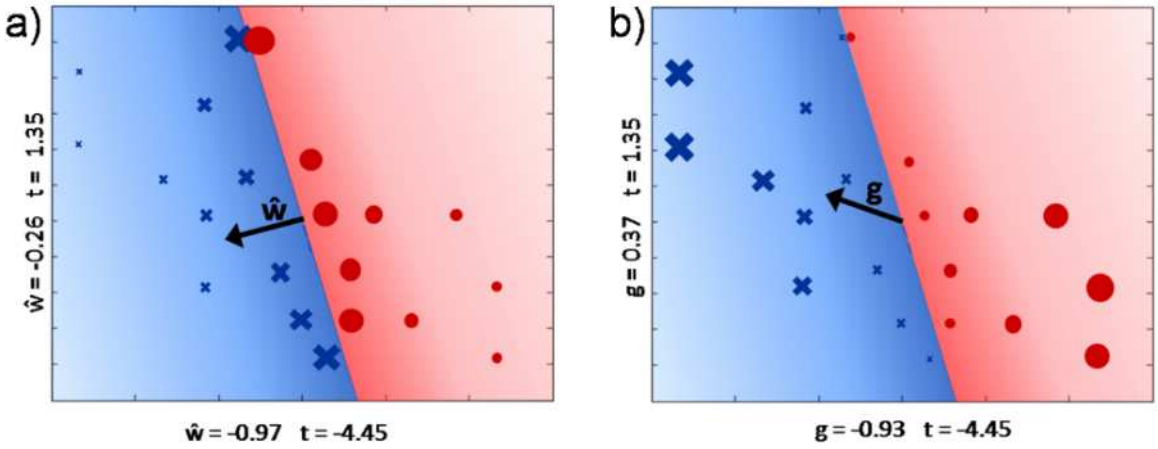

Figure 2. Geometric interpretation of Gaussian process mapping methods in a two-dimensional feature space. Each mapping method constructs a vector from a weighted combination of data points and the size of each data point is proportional to the magnitude of the weighting coefficient. The classifier is trained to separate crosses (class 1) from circles (class 2) and the shaded area indicates distance from the decision boundary. a) In GPC w-mapping, samples closer to the decision boundary carry higher weight and the weight vector ($\mathbf{w}$) is orthogonal to the decision boundary. b) In GPC g-mapping, samples further from the decision boundary carry higher weight and the vector ($\mathbf{g}$) describes how classes are distributed relative to one another. Axis labels contain the projection of the map vector onto each axis and a univariate t-statistic for each data dimension (from Marquand, et al., 2009).

Node-specific group distribution mapping

For single GP classifiers, the two mapping approaches described above are valid representations of the decision boundary and the group distribution, respectively.

However, they provide no information regarding the contribution of a single feature in the context of the combination of multiple classifiers. For the newly developed multi-source classification method, we thus propose an additional mapping algorithm.

The aim is to construct a map quantifying the multivariate group distribution of the two classes for each dimension (feature) from a single biomarker in the context of the optimal tree model (see 2.3.2 Second-level prediction). In other words, we seek to construct a group distribution g-map as described above which is specific to each node of the optimal decision tree. As such, it ought to contrast the two groups of subjects classified at that node considering only the biomarker data for this node.

This is done by constructing a map from the distribution of features at each node of the decision tree, where only the subset of samples found at each node is considered. The weighting factor in this case is the predictive probability adjusted by the node's decision threshold. In analogy to the g-map for single classifiers, it can be calculated by

$$d_{node} = \sum_{i=1}^{m_{node}} (p_i - p_{node})\mathbf{x}_i \qquad \text{equation 8}$$

where m_{node} denotes the number of samples at the respective node, p_i are the GPC predictive probabilities for each sample[10] and p_{node} is the probability threshold for that node determined by the optimal decision tree. The key idea in our approach is to scale the node-specific data x_i with the predictive probabilities p_i for each sample present at the respective node. In order to contrast specifically those subjects which have been classified to different groups at the present node, we set the mean of p_i at each node to 0 by subtracting p_{node}. This assigns those samples classified by the

[10] Note that p_i simply denotes the i^{th} component of the p_{GP} vector for the respective node.

decision tree as being in the right branch a positive weight ($p_i - p_{node} > 0$); those classified as being in the left branch will carry a negative weight ($p_i - p_{node} < 0$).

The map can be interpreted as providing a measure of how samples in the two branches emerging from each node are distributed relative to one another and to the decision boundary provided by the first-level classifier. Thus, to understand which features contribute to prediction within the optimal decision tree, it is necessary to navigate through the tree beginning at the root node. At each node, the contribution of each feature to class distribution is described by equation 8. In analogy to single classifier group distribution g-mapping, the decision boundary is not changed by the procedure, i.e. the hyper-parameters remain the same as for the single classifier mappings. Summarizing, d_{node} -maps show the multivariate, node-specific group distribution of the two classes with respect to the original single classifier decision boundary.

2.4. Summary

In the previous sections, we have outlined how classifiers can predict to which one of two classes (e.g. patient or healthy control) a subject belongs based on high-dimensional data. While recently developed pattern recognition methods are highly successful performing such classifications based on a single high-dimensional biomarker, no currently available algorithm is able to integrate multiple high-dimensional datasets. We addressed this problem by developing a two-level procedure: On the first level, high-dimensional data from each single biomarker is used to predict each subject's probability to be in a certain class. Specifically, we are able to obtain a class probability prediction for each subject and biomarker using a GPC. On the second-level, we then classify each subject based on its GPC class

probability predictions from each biomarker (p_{GP}) using a CART algorithm. In summary, the basic idea consists of first reducing dimensionality of the problem space while not losing essential information on the first level. Then – on the second level – a non-linear CART classifier generates final classifications based on first-level predictions. For the new procedure, we then developed a method enabling significance testing as well as an algorithm allowing for the quantification (mapping) of the contribution of each biomarker's single features to final CART prediction.

In summary, the algorithm now allows for single-subject classification based on multiple high-dimensional biomarkers such as those commonly obtained using neuroimaging or genetic analyses. In addition, it can be assessed which biomarkers are relevant and in what way the relevant markers contribute to classification (e.g. which genes or brain regions play a role for prediction).

3. Part II – Classification in the context of depression

For depression with its high prevalence, high degree of suffering (see 3.1.1.1 Epidemiology of depression), and low treatment efficiency (Hennings, et al., 2009), many attempts to identify biomarkers have been made. To date, however, this research has not yielded biomarkers suitable for practical application (3.2 Summary and goals of the study). As a first step toward improving predictive accuracy in comparison to previous attempts, we will – in *Part II* of this work – suggest combining multiple biomarkers of depression and test the approach by classifying depressive patients and healthy control subjects: First, we will identify the core-symptoms of depression from standard classification systems (3.1.1.2 Symptoms and diagnosis of depression) and outline three experimental paradigms (3.3.2 Tasks and procedures) probing neural processes known to be related to these symptoms (3.1.2 Biological markers of depression). Then, we will use the multi-source pattern classification algorithm developed in *Part I* of this work to integrate these measurements in order to improve classification accuracy. Finally, we will investigate which biomarkers are essential in the integrated biomarker model and outline the neural processes relevant in this context.

3.1. Introduction

In this section, we will introduce the concept of depression, summarize epidemiological data, and outline the symptoms and diagnostic criteria of depression. In order to identify potentially suitable biomarkers, we will then review the literature on biomarkers of depression with a focus on evidence from neuroimaging.

3.1.1. The concept of depression

Colloquially, depression can refer to both a transient mood state as well as to a common affective disorder characterized by persistent negative mood, anhedonia, and deficient cognitive, circadian, and psychomotor functioning (Seminowicz, et al., 2004). In psychiatry, the term is used to denote the latter condition most often called Major Depressive Disorder (for detailed diagnostic criteria and terminology, see 3.1.1.2 Symptoms and diagnosis of depression).[11] Over more than a century, a vast body of theory has grown to explain causes and account for symptoms of the disorder, incorporating advances in the fields of behavioral and cognitive psychology, medicine, and biology. Evidence suggests a complex interaction of personal experience and learning history, individual personality traits, and genetic factors to contribute to affective disorders (Ebmeier, Donaghey, & Steele, 2006; see Gotlib & Hammen, 2009, for an in-depth treatment).

3.1.1.1. Epidemiology of depression

The World Health Organization (WHO) currently considers depression the single most burdensome disease in terms of total disability-adjusted life years for people in mid-life worldwide (Murray & Lopez, 1996). In industrialized nations, it even is the most burdensome disease over all age-groups (Lopez & Murray, 1998). This is due to the interaction of relatively high lifetime prevalence, early age of onset, a substantial rate of chronic manifestation, and high role impairment (Gotlib & Hammen, 2009). In addition to an increased risk of suicide (Bostwick & Pankratz, 2000; Bronisch & Wittchen, 1994), depression has been found to adversely affect

[11] In the context of this work, the term depression (depressive/depressive patients) will consistently be used to refer to individuals who are currently suffering from a Depressive Episode as defined by ICD-10 or DSM-IV or who are recovering from a recent one.

interpersonal relationships (Wade & Cairney, 2000) as well as decrease workplace productivity (Kessler, et al., 2006).

While up to 20% of adults report recent depressive symptoms as assessed by a screening scale (Kessler, et al., 2001), studies using structured diagnostic interviews based on DSM-IV criteria yield point prevalences of Major Depression between 2 and 4% (WHO International Consortium in Psychiatric Epidemiology, 2000). Twelve-month prevalence estimates based on structured diagnostic interviews of adults in the USA are at 6.6%. Importantly, almost 90% of these cases are classified as clinically depressed with moderate, severe, or very severe symptoms according to standard *Hamilton Rating Scale for Depression* severity thresholds (Kessler & Merikangas, 2004). Thus, roughly 5.9% of the population experienced clinically significant depressive symptoms with duration of at least one month during the previous one-year period. Based on the same data, life-time prevalence estimates lie at 16.6% (Kessler, et al., 2005; this result is confirmed by an independent data set by Haro, et al., 2006). With 6.9 and 14%, respectively, twelve-month and life-time prevalences in Europe (Wittchen & Jacobi, 2005) are comparable to rates in the USA.

The onset of depression commonly lies between 19 and 44 years of age (interquartile range; median: 32 years) with 10% of all cases not having the first episode before the age of 55 years (Kessler, et al., 2005). Furthermore, about 80% of depressive patients experience more than one episode in their life (Kessler, et al., 2003). The median recovery time is 6 weeks with 90% recovering within a year (Kendler, Walters, & Kessler, 1997). However, recovery time for psychiatric in-patients is generally longer and more variable (Brugha, et al., 1990).

Epidemiological studies have furthermore identified a number of risk factors for depression, the largest of which are gender and environmental adversities.

Considering life-time prevalence, women are affected roughly twice as frequently as men (Blazer, Kessler, McGonagle, & Swartz, 1994). Additionally, evidence suggests a substantial role of stressful life-events such as unemployment, loss of close personal relationships, and major health problems (Kessler, 1997).

3.1.1.2. Symptoms and diagnosis of depression

Reflecting the fundamental disturbances in affective functioning, depression is generally classified as a mood disorder. Its foremost characteristics are an excessive negative affect (e.g. lowered mood) and a persistent deficiency in positive (e.g. anhedonia). Other symptoms may include alterations in cognitive, circadian, or psychomotor functioning. Most commonly, depression is diagnosed based on the criteria outlined in DSM-IV or ICD-10. Generally, the two standard classification systems define mental disorders based on their symptoms, that is specific observable behavioral and/or cognitive patterns, rather than on etiology or pathophysiological mechanisms.

In DSM-IV, *Major Depression* is characterized by the occurrence of one or more depressive episodes. Diagnosis of such an episode requires at least five out of the following nine symptoms to be present during the same 2-week interval:

1) depressed mood on almost all days for most of the day,
2) diminished interest or pleasure in all or almost all activities on almost all days for most of the day,
3) weight gain or loss (without a diet) or change of appetite,
4) insomnia or hypersomnia,
5) psychomotor agitation or retardation,
6) fatigue or loss of energy,

7) feelings of worthlessness or excessive or inappropriate guilt,

8) diminished ability to think or concentrate or indecisiveness,

9) recurrent thoughts of death or suicidal ideation or suicide attempt.

Importantly, the two core-symptoms – depressed mood and diminished interest or pleasure – have to be present among these five symptoms.

In its section on *Mood Disorders*, the ICD-10 distinguishes between the concepts of *Depressive episode* and Recurrent Depressive Disorder. The latter is defined by the occurrence of at least two *Depressive episodes*. In analogy to DSM-IV criteria, lowered mood and a reduced capacity for enjoyment and interest are the essential symptoms with lowered mood varying little from day to day, independent of circumstances. The additional symptoms also virtually mirror the ones listed in DSM-IV (for a systematic comparison see Gruenberg, Goldstein, & Pincus, 2005).

In summary, both standard classification systems ground their definitions in the occurrence of at least one depressive episode. This episode, in turn, necessarily comprises symptoms of lowered mood and anhedonia while other symptoms are non-essential for diagnosis.

3.1.2. Biological markers of depression

Over more than two decades, research has consistently identified differences within the nervous system between depressive patients and healthy controls. It is beyond the scope of this work to provide a comprehensive review of the vast literature on these differences. As affective alterations are most essential in depression – both in terms of diagnosis (3.1.1.2 Symptoms and diagnosis of depression) and individually perceived degree of suffering – we will focus on behavioral and neurophysiological deviations related to these symptoms, while

omitting the field of cognitive deficits and their physiological correlates (for a summary, see Mössner, et al., 2007). In the following, we will first briefly outline deviations in emotional processing as they relate to the cardinal symptoms of depression to then focus on the neural basis of depression before closing with a concise summary of other biomarkers which have been suggested within the field of neuroscience.

3.1.2.1. Processing of emotional stimuli

The two cardinal symptoms of depression – lowered mood and anhedonia – are related to a number of altered affective and motivational processes which affect behavior and cognition of depressed individuals (for theoretical consideration on how mood might affect behavior/cognition, see Rottenberg & Johnson, 2007). As the identification of neurobiological markers – particularly those derived from functional neuroimaging – is based on neural processes associated with these altered cognitions and behaviors, we will, in the following, outline key-findings in this area.

Studies investigating the processing of emotional stimuli have focused mainly on the investigation of emotional facial expressions (for reviews, see Leppanen, 2006 and Bylsma, Morris, & Rottenberg, 2008). In this line of research, patients suffering from major depressive disorders showed impairments in the identification of affect in happy and sad facial expressions (Rubinow & Post, 1992; Surguladze, et al., 2004). Furthermore, patients were shown to preferentially attend to sad facial expressions: In a dot-probe task, subjects are simultaneously shown two photographs depicting faces with different emotional expressions – one on the left and one on the right side of the screen. After the disappearance of the pictures, a dot (probe) is shown in the location of one of the previously seen photographs (cues) and subjects are instructed

to indicate the position of the dot (left or right) by pressing one of two buttons. At a delay of 1000 ms between the cues and the probe, subjects suffering from Major Depression respond significantly faster if the probe is located in the position of the previously displayed sad facial expression. This effect is not seen for neutral, happy, or angry facial expressions. From this, the authors conclude that clinically depressed individuals show an attentional bias towards sad emotional stimuli. Importantly, this effect was not found in patients suffering from Generalized Anxiety Disorder or Social Phobia suggesting that the attentional bias to sad faces might be specific to depression (Gotlib, Krasnoperova, Yue, & Joormann, 2004; Gotlib, et al., 2004a).

Surguladze et al. (2004) investigated happy facial expressions: Comparing healthy controls with depressed patients, they demonstrated impairments in discrimination accuracy for mildly happy expressions. However, Suslow et al. (2004) showed a similar bias away from happy faces only for those depressive patients suffering from comorbid anxiety while they failed to find the effect in a sample of depressive patients who were free of comorbidities. Other studies raise similar concerns related to the specificity of the effect to depression (Gilboa-Schechtman, Presburger, Marom, & Hermesh, 2005; Gotlib, Kasch, et al., 2004).

Also, the processing of neutral facial expressions is altered in acutely depressed individuals: When asked to indicate whether a facial expression is neutral, happy, or sad, depressive persons – in comparison to healthy controls – falsely classify significantly more neutral expressions as sad. Also, the time needed for classification of neutral faces is higher in depressive individuals (Leppanen, Milders, Bell, Terriere, & Hietanen, 2004). From this, it appears that depressive individuals do not perceive neutral facial expressions as unambiguously neutral, but need more time to process the stimuli while still erroneously attributing sadness to them.

In addition to these findings concerning mainly the online processing of emotional facial expressions, depressive patients have consistently shown a memory bias toward mood congruent stimuli. Specifically, they remember more items with negative (e.g. words such as “guilt” or “tears”) than with positive or neutral valence in free recall tasks. This is true also for supra- as well as subthreshold recognition tasks (for a review, see Colombel, 2007).

Moreover, evidence shows alterations in the responsiveness to reward and punishment: Henriques & Davidson (2000) conducted a verbal memory task in which subjects suffering from depression failed to alter their strategy of responding in reaction to changing monetary reward. However, their response changed in reaction to monetary loss, demonstrating an understanding of the task, as well as an ability to generally strategically adapt response patterns. This effect was still present after exclusion of subjects with comorbid symptoms of anxiety (compare potentially contradicting evidence related to the attentional bias away from possibly rewarding happy faces). The same group had previously shown this effect in a non-clinical sample showing depressive symptoms (Henriques, Glowacki, & Davidson, 1994). Furthermore, an increased sensitivity to failure (Elliott, et al., 1996) has been observed. Specifically, depressive patients – in comparison to healthy controls – were much more likely to commit an error in a series of neuropsychological test items, if they had failed to solve the previous item.

In summary, depressive patients show an increased propensity to negative emotional processing as well as altered reward processing (for reviews see Leppanen, 2006; Bylsma, et al., 2008; Chau, et al., 2004; Drevets, 2001).

3.1.2.2. Neuroimaging markers

Anatomical and resting state studies

Converging evidence has been accumulated which suggests the involvement of a wide network of distributed brain structures in the pathological processes relevant in depression. Using Positron Emission Tomography (PET) during wakeful rest, deviating cerebral blood flow and glucose metabolism within the prefrontal cortex and the limbic system have consistently been found in depressive patients in contrast to healthy controls. Studies report reduced metabolism mainly in the anterior cingulate cortex as well as in frontal regions. Elevated activation is consistently found in the amygdala, medial thalamus, and ventrolateral and orbitofrontal areas (Ito, et al., 1996; Kennedy, Javanmard, & Vaccarino, 1997; for reviews see Drevets, 2001; Manji, Drevets, & Charney, 2001). Moreover, glucose metabolism in limbic, thalamic, and basal ganglia structures predicts symptom severity as measured with the *Hamilton Depression Rating Scale* (HDRS, Hamilton, 1960) further elucidating the involvement of these regions in depression.

In an attempt to identify regions specifically associated with psychopathological components, three HDRS factors were found to correlate significantly with glucose metabolism: Psychic depression – mainly mirroring depressed mood, suicidal ideations, and feelings of worthlessness and hopelessness – correlated positively with metabolism in the cingulate gyrus, thalamus, and basal ganglia. Sleep disturbance correlated positively with metabolism in limbic structures and basal ganglia while loss of motivated behavior was negatively associated with parietal and superior frontal regions (Milak, et al., 2005).

Studies employing structural neuroimaging methods have shown reductions in volume for the orbitofrontal cortex (Bremner, et al., 2002), the subgenual anterior cingulate gyrus (Botteron, Raichle, Drevets, Heath, & Todd, 2002; Drevets, et al.,

1997), the amygdala (Sheline, Gado, & Price, 1998), and the hippocampus (Bremner, et al., 2000; Sheline, Sanghavi, Mintun, & Gado, 1999) in depressive patients. Additionally, reward-related structures such as the putamen (Husain, et al., 1991), and the caudate (Krishnan, et al., 1992) displayed decreases in volume. Likewise, drug-naïve depressive patients suffering from their first major depressive episode showed significant gray-matter volume reduction in limbic regions including hippocampus and parahippocampus extending into the medial temporal lobe (Zou, et al., 2010).

In addition, research into the causes of differential treatment response has provided particularly valuable insights (for a review, see MacQueen, 2009): Using a multivariate pattern recognition algorithm on anatomical MRI data, Costafreda et al. (2009) were able to predict response to pharmacological treatment (but not to cognitive behavioral therapy) with an accuracy of 89%. In particular, increased grey matter density in the anterior and posterior cingulate cortices increased the probability of clinical remission in response to fluoxetine. Increased density in the orbitofrontal cortex increased the probability of residual symptoms of depression following this medication. With the same approach, diagnostic classification of depressive patients and controls was, however, only at 68% (Costafreda, et al., 2009). Further evidence additionally suggests regional cerebral blood flow in the rostral anterior cingulate to predict response to medication (Joe, et al., 2006; Mayberg, et al., 1997). Using electrophysiological methods, this result was replicated (Mulert, et al., 2007) and extended by Pizzagalli et al. (2001) who showed that activation in the rostral part of the anterior cingulate is associated with magnitude of response to pharmacological treatment. To this end, relative theta power of the EEG measured during the first week of treatment predicted response to selective serotonin (5-HT) reuptake inhibitors (SSRIs; Iosifescu, et al., 2009).

In this context, the fact that depression can be treated effectively with SSRIs gave rise to the notion that 5-HT plays a decisive role in major depressive disorder. The principal centers for serotonergic neurons are the rostral and caudal raphe nuclei. From the rostral raphe nuclei, axons ascend to the cerebral cortex, limbic regions and specifically to the basal ganglia. Serotonergic nuclei give rise to descending axons, some of which terminate in the medulla, while others descend the spinal cord. Generally, depression is associated with alterations in 5-HT neurotransmission and studies show that tryptophan depletion can induce symptoms of depression (Delgado, et al., 1990): P-chlorophenylalanine (PCPA), a specific inhibitor of tryptophan hydroxylase which is the rate limiting enzyme in the biosynthesis of 5-HT, and a tryptophan-free drink induced a rapid onset of clinical depression in patients with previous depressive episodes. Furthermore, in patients with a depletion-induced depressive relapse, tryptophan depletion resulted in a decreased glucose metabolism in the middle frontal gyrus, thalamus, and orbitofrontal cortex. Validating these results, decreased glucose metabolism in these regions correlated with increased depressive symptoms (Bremner, et al., 1997). Within this framework, Pezawas et al. (2005) reported effects of *5-HTTLPR* genotype on grey matter volume in regions relevant for depression, namely perigenual cingulate and amygdala. Directly linking depression to 5-HT neurotransmission, Reivich et al. (2004) showed increased serotonin transporter availability in frontal and cingulate regions using a radioligand which binds specifically to the 5-HT transporter. Using a similar approach, decreased 5-HT1A receptor levels in depressive patients in midbrain raphe, parietal, and occipital cortex regions have been shown (Drevets, et al., 2000). While this effect cannot be shown in midbrain raphe, cortical areas continue to display lowered 5-HT1A receptor levels after recovery (Bhagwagar, et al., 2004).

Moreover, neuroimaging findings as well as evidence from animal models strongly suggest an involvement of the dopaminergic system in depression (for reviews, see Dunlop & Nemeroff, 2007; Nestler & Carlezon, 2006). In this line of research, the mesolimbic dopamine pathway linking the ventral tegmentum to the ventral striatum, the hippocampus, the amygdala, and the septum, has commonly been associated with reward processing (Day & Carelli, 2007). Thus, especially anhedonia as one of the major symptoms of depression has been the focus of attention: Tremblay et al. (2005) were able to verify the involvement of dopamine-related neuroanatomical structures in altered reward processing in major depressive disorder: Depressive patients showed a hypersensitivity to the rewarding effects of dextroamphetamine associated with altered brain activation in ventrolateral prefrontal cortex, orbitofrontal cortex, caudate, and putamen. In addition, functional polymorphisms of the D_4 receptor, dopamine transporter, and Catechol-O-methyl transferase, all closely associated with dopamine metabolism, have been shown to influence depression-related processes (for a review, see Dunlop and Nemeroff, 2007). In accordance with Davidson and colleagues' (2002) considerations concerning the heterogeneity of symptoms clustered to form a DSM-IV or ICD-10 diagnosis, Bragulat et al. (2007) found that different symptoms (affective flattening, psychomotor retardation, and impulsivity) in depression entail different regional presynaptic dopaminergic function in the caudate, parahippocampus, and parahippocampal gyrus.

In addition to the roles of 5-HT and dopamine, γ-aminobutyric acid (GABA) neurotransmission has recently drawn increasing interest: Studies utilizing Magnetic Resonance Spectroscopy (MRS) revealed decreased frontal and occipital GABA concentrations while glutamate concentrations in occipital regions were elevated in depressive patients compared to controls (Sanacora, et al., 2004; Sanacora, et al., 1999; Hasler, et al., 2007). Furthermore, evidence suggests that decreased GABA

and increased glutamate levels in frontal and occipital-parietal regions persist after recovery (Bhagwagar, et al., 2007; Bhagwagar, et al., 2008). It is speculated that lowered glial cell numbers which have been reported in affective disorders (Harrison, 2002; Ongur, et al., 1998) might cause decreased GABA levels as particularly astrocytes are an essential source of the GABA precursor glutamine (Bhagwagar, et al., 2008).

Imaging of task-related activation

The analysis of task-related activation is of particular interest as it allows for a specific probing of neural processes during behavioral tasks. Focusing on the processing of emotional facial expressions, neural activity in depressive patients, but not in controls, was found to increase linearly in response to increasingly intense sad faces in areas known to be involved in emotional processing (putamen, parahippocampal gyrus/amygdala) and in the analysis of stimulus features (fusiform gyrus). In response to increasingly intense happy faces, a linear increase in putamen and fusiform gyrus was observed in healthy controls, but not in patients (Surguladze, et al., 2005). Evidence from a study employing multivariate pattern classification showed substantial contributions of frontal regions (middle and superior frontal gyrus) as well as precuneus, postcentral gyrus, inferior occipital gyrus, and fusiform and lingual gyri to the correct classification of depressive patients based on the neural response during the presentation of sad faces (accuracy = 74% and 76% for medium and high intensity sad faces, respectively). Corresponding to the impaired recognition of neutral facial expressions on the behavioral level (Leppanen, et al., 2004), depressive patients could also be identified based on their neural response pattern following neutral facial expressions (accuracy = 87%). Here, again frontal regions as well as precuneus, postcentral gyrus, inferior occipital gyrus, and lingual gyrus

contributed to classification accuracy. In addition, involvement of the anterior cingulate and the parahippocampal gyrus was shown (Fu, et al., 2008).

In response to negative words, Siegle et al. (2002) found prolonged activation in the amygdala. While activation in healthy controls decayed within 10 seconds, activation in depressive patients continued for roughly 30 seconds. In particular, it also continued when a distracting task (Sternberg memory paradigm) followed the negative word.

As for regional cerebral blood flow during rest (see above), evidence suggests that treatment response can also be predicted from task-related functional neuroimaging data (Siegle, Carter, & Thase, 2006): Specifically, response to cognitive behavioral therapy was optimal if patients' sustained reactivity to emotional stimuli was low in subgenual cingulate cortex while activation of the amygdala was high. In the same line of research, treatment response was predicted with more than 78% accuracy using multivariate pattern recognition algorithms based on neural responses following the presentation of sad facial expressions. Regions contributing most to classification included a complex network of anterior cingulate, frontal regions, as well as occipital and parietal areas. Interestingly, using neutral facial expressions with the same algorithm provided an equally large accuracy rate (Costafreda, et al., 2009).

In addition, neuroimaging studies integrating data from genetic analyses provide detailed information on the role of 5-HT: Heinz et al. (2005) provided evidence that a polymorphism of the serotonin transporter gene (SLC6A4) modulates amygdala-prefrontal coupling. Moreover, Pezawas et al. (2005) performed a functional analysis of perigenual cingulate and amygdala – for which volumetric differences associated with a functional 5' promoter polymorphism of the serotonin transporter gene have been shown (see above) – demonstrating differences in coupling during perceptual

processing of fearful stimuli. Specifically, short-allele carriers showed relative uncoupling of this circuit.

Focusing specifically on the neural correlates of anhedonia in depression, Keedwell et al. (2005) showed that severity of anhedonia is positively correlated with reward-related activity in the ventral medial prefrontal cortex while it is negatively correlated with activation in the amygdala and the ventral striatum. In line with this, a decreased response of ventral striatal structures to rewards has consistently been observed in depressive patients (Epstein, et al., 2006; Pizzagalli, et al., 2009). During anticipation of rewards, patients displayed increasing anterior cingulate activation with increasing magnitude of reward (Knutson, Bhanji, Cooney, Atlas, & Gotlib, 2008).

3.1.2.3. Other biological markers

In addition to the behavioral and neuroimaging biomarkers outlined above, a large number of candidate biomarkers have been suggested from various fields of neuroscience and medicine (for a review, see Mössner, et al., 2007). In the following, we will summarize a number of the more robust findings.

One of the most reliable markers for depression is a decreased imipramine binding to the high-affinity serotonin transporter on platelets: Meltzer and Arora (1986) found a generally decreased binding capacity as well as a decreased binding affinity in depressed subjects. This is supported by a meta-analysis of 76 studies (Ellis & Salmond, 1994). These findings also hold if only high-affinity binding sites are considered (this is of particular importance, as results considering low-affinity binding sites have been methodologically criticized; Mellerup & Plenge, 1988). While arguing for the face validity of the results, the fact that the decrease in binding capacity was

substantially reduced in medicated patients compared to subjects who had been medication-free for more than four weeks, might pose difficulties for diagnostic application.

Investigating the serum, increased levels of interleukin (IL)-6 and soluble IL-2 receptor (sIL-2R; Maes, Meltzer, Bosmans, et al., 1995; Maes, Meltzer, Buckley, & Bosmans, 1995; Sluzewska, et al., 1996) have been observed in depressive patients. In the context of classification, it appears noteworthy that Maes et al. (1995) used Linear Discriminant Analysis (LDA) with IL-6, sIL-6R, sIL-2R, and the highly correlated transferrin receptor as well as the product term IL-6 * sIL-6R obtaining a within-sample accuracy of 84%. Unfortunately, the authors did not test the LDA function on independent data (i.e. cross-validate their results) to obtain an estimate of the accuracy within the population. Thus, even neglecting the problem of overfitting with four intercorrelated parameters, this value must be seen as the upper bound of accuracy that can possibly be obtained using the parameters.

Additionally, serum levels of brain derived neurotrophic factor (BDNF; Karege, et al., 2002) are decreased while several fibroplast growth factor (FGF) system transcripts have been found to be disregulated in depression (Evans, et al., 2004; for a review of these and other markers, see Mössner, et al., 2007). Both neurotrophins might be of particular value in the search for treatment response markers as both substances have been associated with the mechanism of action of SSRIs (Russo-Neustadt, et al., 2001; Evans, et al., 2004). Moreover, FGF gene expression is altered in unipolar depressive individuals, but not in bipolar patients stressing the specific diagnostic potential of this parameter (Evans, et al., 2004).

3.2. Summary and goals of the study

Summary

Depression is a highly common, in many cases severe and often chronic disorder with significant consequences for subjective quality of life, interpersonal relationships, and working productivity. While substantial and persistent lowered mood and anhedonia have to be present to justify the diagnosis of depression, the frequently encountered impairments of cognitive, circadian, and psychomotor functioning are non-essential. Furthermore, research has provided compelling evidence showing altered affective and motivational processing on the behavioral level while in-vivo neuroimaging studies have begun to elucidate the complex neural underpinnings of pathological deviations in depression. Generally, the involvement of prefrontal and occipital regions, the basal ganglia, and the limbic system in depression has reliably been replicated using different techniques and paradigms.

Based on this evidence, substantial attempts have been made to identify potential biomarkers of depression which might aid in diagnosis or predict treatment response or course of illness. However, data on single subject classification accuracy is particularly scarce: For instance, of the 31 studies considered above which directly compare depressive patients and healthy controls, the majority (19) do not provide sufficient information for accuracy calculations (mainly as a single effect size estimation is not feasible for mass univariate neuroimaging analyses). However, two studies do provide direct accuracy estimates while for another ten studies it was possible to calculate accuracies from statistical data given in the respective articles: Generally, accuracies range between .56 and .87 (Table 1. Accuracy estimates of previously identified biomarkers of depression).

Table 1. Accuracy estimates of previously identified biomarkers of depression

Biomarker description	accuracy estimate[12]	study reference
bias attending to sad facial expressions	< .57	Gotlib, et al., 2004
identification of neutral facial expressions (hit-rate)	< .70	Leppanen, et al., 2004
lower occipital GABA concentration	< .65	Sanacora, et al., 1999
	< .75	Sanacora, et al., 2004
lower frontal GABA concentration	< .57	Hasler, et al., 2007
lower occipital glutamate concentration	< .80	Sanacora, et al., 2004
amygdala activation following negative words (left; right)	< .78; < .76	Siegle, et al., 2002
structural whole-brain MRI	.68	Costafreda, et al., 2009
whole-brain fMRI BOLD response to neutral faces	.87	Fu, et al., 2008
whole-brain fMRI BOLD response to mildly sad faces	.74	Fu, et al., 2008
whole-brain fMRI BOLD response to highly sad faces	.76	Fu, et al., 2008
imipramine binding to platelets	< .61	Meltzer & Arora, 1986
sIL-2R serum levels	< .67	Maes, et al., 1995
	< .78	Sluzewska, et al., 1996
BDNF serum levels	< .56	Karege, et al., 2002

From these mixed results it appears that among the five studies obtaining the highest accuracy, four are functional neuroimaging studies which show deviations in task-related activation – three use multivariate methods. Against this background, it

[12] For those studies which provided a single statistical parameter (e.g. t- or F-values) meaningful for group comparison, Pearson correlation r was calculated as a measure of effect size (respective formulas are provided in Hunter & Schmidt, 1990). Then accuracy was calculated using the Binomial Effect Size Display principle (Rosenthal & Rubin, 1982; assuming a symmetrical cost function it follows: $accuracy = \frac{r}{2} + \frac{1}{2}$).

is essential to note that none but the two studies which provided direct accuracy estimates tested their results on independent data as no single-subject prediction was originally intended. As they, thus, did not obtain an unbiased estimate of the accuracy within the population, the values represent the upper bound of the accuracy that can optimally be obtained using the studies' parameters.[13] In addition, the real-life utility of the study which obtained the by far highest accuracies on independent data (up to 87%; Fu, et al., 2008) cannot be assessed easily: Primarily, the use of a sample containing a group of unmedicated and acutely depressive patients makes an estimation of the accuracy in a heterogeneous psychiatric sample difficult. Considering these limitations, it must be concluded that – while using a multivariate method with task-related functional neuroimaging data appears most promising – none of the currently available biomarkers has shown sufficient predictive power for practical application in psychiatry.

Goals of the study

In order to improve the predictive power obtained with biomarkers, it might be most instructive to consider the diagnostic process employed in psychiatry today: Diagnoses made by experienced clinicians using standardized tools are undoubtedly the gold standard on which decision-making in psychiatry is based.[14] As psychiatric disorders are defined by specific behavioral and cognitive patterns, the clinician needs to integrate his observations and the patient's reports for each symptom to obtain reliable and valid data based on which the presence or absence of a defined

[13] For descriptive reasons, we ran the decision tree algorithm without using independent data (i.e. without leave-one-out cross-validation) and obtained an accuracy of 95%. Running single GP classifiers without cross-validation usually yields perfect accuracies due to the ultra-high dimensionality of the data.

[14] Note that accuracy in primary care settings differs: General practitioners classify about 35% of their patients incorrectly (accuracy = .65; not depressed/depressed) while not detecting 51% of depressive patients (sensitivity = .49; Mitchell, Vaze, & Rao, 2009).

disorder is judged. Following the same principle, integrating multiple physiological markers tied to specific symptoms ought to improve both validity and accuracy of biomarker-based classification.

With its ability to assess neural responses directly linked to behavior and cognition (see 3.3.3 Functional Magnetic Resonance Imaging), fMRI can provide data on single symptom-related processes. As outlined above, neural correlates of specific behaviors and cognitions have already been identified and used successfully in classification (3.1.2 Biological markers of depression). Using the multi-source pattern classification algorithm developed in this work, an integration of such symptom-related markers measured with fMRI ought to be possible. Considering previous results and the prominent role of lowered mood and anhedonia in depression as defined by DSM-IV and ICD-10, combining neuroimaging data measured during emotional processing appears most promising. Specifically, we will measure neural responses to four different emotional facial expressions as well as to reward- and loss-related stimuli of varying intensity (3.3.2 Tasks and procedures). Thereby, we incorporate neural responses associated with behavioral and cognitive patterns related to the core-symptoms of the disorder which have robustly been shown to deviate in depression.

In order to provide a realistic estimate of the approach's potential real-life utility, it is essential to run the algorithm on a sample of patients as it can commonly be found in psychiatric settings. This requires inclusion of a heterogeneous group of patients regardless of current therapeutic intervention or medication who present with varying degrees of depressive symptoms.

In summary, we will use the multi-source pattern classification algorithm developed in this work to integrate fMRI BOLD data acquired with multiple paradigms and conditions related to emotional processing and anhedonia. Following our

approach, we will use the GP classifiers to obtain a participant's probability of being a patient for each of the symptom-related neural processes. In the second step, these classification probabilities associated with each biomarker will be integrated using a decision tree algorithm (compare 2.3 Algorithm development).

We hypothesize that

(1) single GP classifiers based on neural correlates of the processing of emotional facial expressions can classify depressive patients and controls which have not previously been seen by the algorithm with significant accuracy.

(2) single GP classifiers based on neural correlates of the processing of reward- and loss-related stimuli can classify depressive patients and controls which have not previously been seen by the algorithm with significant accuracy.

(3) combining the predictive probabilities obtained from each GP classifier will result in significantly increased classification accuracy compared to the accuracy obtained from the most accurate of the single GP classifiers alone.

If high-accuracy classification is possible, it is of great interest how classification was achieved and which properties of the data have contributed. Thus, we will quantify the utility of each single biomarker and derive a decision tree which models the interrelations of the relevant markers. Using multivariate spatial mapping, we will furthermore identify those brain regions which contributed most to overall classification within the decision tree. Finally, we will discuss the resulting integrated biomarker model and the network of contributing regions in the context of depression.

3.3. Materials and Methods

3.3.1. Participants

A total of 31 psychiatric in-patients from the University Hospital of Psychiatry, Psychosomatics and Psychotherapy (Würzburg, Germany) were recruited. All were in a depressed phase or recovering from a recent one. Specifically, patients were diagnosed with recurrent depressive disorder (F33; n=10), depressive episodes (F32; n=15), or bipolar affective disorder (F31; n=5) based on the consensus of two trained psychiatrists according to ICD-10 criteria (DSM-IV codes 296.xx). In addition, patients were interviewed using the Montgomery-Åsberg Depression Rating Scale (MADRS; Montgomery & Asberg, 1979). One patient was excluded due to a panic attack during the measurement, leaving 30 patients for further analysis.

We explicitly recruited patients who were on a variety of medications and who, at the time of the measurements, presented with varying degrees of depressive symptoms from severe to currently almost symptom-free. Accordingly, self-report scores in the German version of the *Beck Depression Inventory – Second Edition* (BDI-II; Beck, et al., 1996) on the day of the experiment ranged from 2 to 42 (mean = 19, standard deviation SD = 9.4). Choosing a well-diagnosed, but heterogeneous group of patients with varying degrees of depressive symptoms while not excluding medicated patients ought to provide a more realistic estimate of the algorithm's potential real-life utility. Exclusion criteria were age below 18 or above 60 years, comorbidity with other currently present Axis I disorders, mental retardation or mood disorder secondary to substance abuse, medical conditions or medication as well as severe somatic or neurological diseases. Patients suffering from bipolar affective disorder were in a depressed phase or recovering from a recent one with none showing manic symptoms. All patients were taking standard antidepressant

medications, consisting of SSRIs, tricyclic antidepressants, selective noradrenaline re-uptake inhibitors, noradrenaline and serotonin selective inhibitors, or 5-HT2 antagonists. Patients were free of antipsychotic medication with the exception of quetiapine for which up to 300 mg/day were generally allowed. However, no dose of quetiapine was given on the day of the experiment so that at least 16 hours (≈ 3-4 plasma half-life times and ≈ 2-3 D_2-receptor half-life times; Gefvert, et al., 2001) had passed since the last administration.

Thirty comparison subjects from a pool of 94 participants previously recruited by advertisement from the local community were selected as to match the patient group in regard to gender, age, smoking, and handedness using the *optimal matching* algorithm implemented in the MatchIt package (Ho, Imai, King, & Stuart, 2007) for R (http://www.r-project.org/). For a summary of the demographic features of the matched groups see Table 2.

Table 2. Demographic features of the matched samples

Variable	Patients n=30	controls n=30
Gender (male/female)	18/12	19/11
mean age (SD)	38.1 (11.0)	36.0 (9.1)
Smoking status (smokers/non-smokers)	14/16	12/18
handedness (right/left)	28/2	29/1
BDI-II score (SD)	19 (9.4)	4 (4.6)

In order to exclude potential Axis I disorders, the German version of the Structured Clinical Interview for DSM-IV (SCID; Wittchen, Zaudig, & Fydrich, 1997) Screening Questionnaire was conducted. Additionally, none of the control subjects showed pathological BDI-II scores (mean = 4.3, SD = 4.6).

From all participants, written informed consent was obtained after complete description of the study. The study was approved by the Ethics Committee of the University of Würzburg and all procedures involved were in accordance with the latest version (fifth revision) of the Declaration of Helsinki.

3.3.2. Tasks and procedures

Setting and external conditions

All measurements were conducted at the Research Center Magnetic Resonance Bavaria (MRB e.V., Würzburg, Germany). While the control subjects independently arranged transportation, patients were met at their ward and taken to the site under supervision of a member of the examination team (psychologist or physician). At the MRB, the details of the study protocol were explained and participants completed the German version of the *Beck Depression Inventory – Second Edition* (BDI-II; Beck, et al., 1996; German version: Hautzinger, Keller, & Kühner, 2006). For other studies, participants completed five additional psychometric paper-pencil tests, namely the *Sensitivity to Punishment Sensitivity to Reward Questionnaire* (SPSRQ; Torrubia, Ávila, Moltó, & Caseras, 2001), the *Zahlen-Verbindungs-Test* (ZVT; Oswald & Roth, 1987), the Positive and Negative Affect Schedule in its state version (PANAS; Krohne, Egloff, Kohlmann, & Tausch, 1996), State-Trait-Angstinventar (STAI; Laux, Glanzmann, Schaffner, & Spielberger, 1981) and State-Trait-Ärgerausdrucks-Inventar (STAXI; Schwenkmezger, Hodapp, & Spielberger, 1992). Controls furthermore completed the Anxiety Sensitivity Index (ASI; Alpers & Pauli, 2001), the *Panik und Agoraphobie-Skala* (PAS; Bandelow, 1997) and the German version of the Structured Clinical Interview for DSM-IV (SCID; Wittchen, et al., 1997) Screening Questionnaire. Then, standardized instructions for all three paradigms were given

and potential questions were addressed. For safety reasons, all participants were screened for metallic objects directly before entering the MRI chamber.

In the MRI measurement room, participants were acquainted with the response pad needed during the second and third paradigm. Subsequently, participants were ear-plugged, comfortably placed on the stretcher, and moved into the scanner. Head movements were minimized by using a cushioned head fixation device. Stimuli were presented via MRI-compatible goggles (VisuaStim; Magnetic Resonance Technologies, Northridge, CA). Even though participants were instructed to lie still and focus on the task, communication was possible at all times via a microphone and speakers in the MRI chamber. Additionally, an emergency button was put on the subject's chest. After completion of the three paradigms, control subjects participated in an additional experiment. Furthermore, anatomical and resting state MRI data was acquired from all participants after the functional paradigms. In total, patients and controls spent no more than 45 minutes in the scanner. After the study, participants were debriefed, blood for genetic analysis required in another study was taken from the control subjects, and patients were escorted back to their ward. Blood from the patients was collected in their ward.

Task description

FMRI data was acquired during a total of three independent paradigms: The first paradigm consisted of passively viewing emotional faces. Sad, happy, anxious, and neutral facial expressions were used in a blocked design, with each block containing faces from eight individuals (four female, four male) that were taken from the Karolinska Directed Emotional Faces database (Lundqvist, Flykt, & Öhman, 1998). Every block was repeated four times in a random fashion. Each face was shown against a black background for two seconds and was directly followed by the next

face. Thus, each block had a duration of 16 seconds. Face blocks were alternated with blocks of the same length showing a white fixation cross on which the participant had to focus. Subjects were instructed to attend to the faces and empathize with the emotional expression.

The second paradigm was a modified version of the Monetary Incentive Delay (MID) Task developed by Knutson et al. (2001) which has been used previously (Hahn, et al., 2009). In order to familiarize subjects with the task, each participant completed ten practice trials prior to data acquisition. During each trial, participants saw one of three different cue shapes (presentation time 2000 ms each) followed by a fixation cross as they waited a variable interval (2250 – 2750 ms). Thereafter, they had to respond in-time (i.e. while the target was visible) with a button press to a white target square which appeared for a variable length of time depending on the subject's previous performance. Specifically, the mean reaction time obtained from the ten practice trials was used as the initial target duration. It was increased by 30 ms if the subject failed to respond fast enough on more than one out of the last three consecutive trials. Likewise it was decreased by 30 ms if the subject succeeded on more than two out of the last three consecutive trials. This approach ought to ensure participants' success on an average of 66% of the trials, thereby yielding a proportion of hits and misses comparable to that reported by Knutson et al. (2001). Additionally, target duration was set as to never decrease below 100 ms and never exceed 1000 ms. As this adaptive algorithm was used to alter target durations, reaction times cannot be meaningfully interpreted and are therefore excluded from further analysis. Feedback (2000 ms), which followed the disappearance of the target, informed participants of whether they had reacted in time during that trial and indicated their cumulative total win in Euros at that point. Cues signaled the possibility of winning 0.05 € (n = 20; a circle with one horizontal line) or 1.00 € (n = 20; a circle with three

horizontal lines). The third cue (n = 20; a triangle) indicated that no money could be won during this trial. The three trial types were randomly ordered within the experiment and the length of the inter-trial interval was randomly jittered in steps of 83 ms between 83 and 2000 ms.

The third paradigm was also adapted from Knutson et al. (2001) and exactly mirrored the second paradigm. However, participants started with an initial amount of 10 Euros of which they were instructed to lose as little as possible. In contrast to the second paradigm, cues signaled the possibility of losing 0.05 € (n = 20; a square with one horizontal line) or 1.00 € (n = 20; a square with three horizontal lines). The third cue (n = 20; a triangle), again, indicated that no money could be lost during this trial.

3.3.3. Functional Magnetic Resonance Imaging

Blood Oxygen Level Dependent (BOLD) imaging

Generally, fMRI relies on measuring the hemodynamic response as a surrogate of neural activity. Specifically, fMRI BOLD measurement takes advantage of the fact that changes in neural activation lead to regional changes in the concentration of oxygenated (O_2Hb) and deoxygenated hemoglobin (HHb) by means of neurovascular coupling: Following neural firing, regional O_2Hb levels decrease as neurons use oxygen, thereby increasing the relative level of HHb in the blood (Heeger & Ress, 2002; Vanzetta & Grinvald, 1999). Following this initial, often very light effect, a much larger increase of O_2Hb levels occurs due to a massive oversupply of oxygen-rich blood. O_2Hb levels reach their maximum after approximately 6 seconds (Fox, Raichle, Mintun, & Dence, 1988; Heeger & Ress, 2002). The result of this oversupply of oxygen is a large decrease in the relative level of HHb. Eventually, the level of HHb slowly returns to its original baseline level after an initial undershoot after

approximately 24 seconds (Heeger & Ress, 2002). As O_2Hb is diamagnetic while HHb is paramagnetic, the relative level of HHb can be assessed with fMRI (for details concerning MRI physics underlying BOLD imaging, see e.g. Filippi, 2009).

It must be noted that assessing the concentration changes of O_2Hb and HHb in the brain is an indirect measure of neural activity, as outlined above. This entails that any event leading to a vascular response in the brain leads to signal changes in the fMRI BOLD raw data. Moreover, irregularities in neurovascular coupling as associated with disorders impacting neurovascular processes (Iadecola, 2004) might also hamper interpretation of BOLD signal changes. However, by means of event-locked extraction and modeling procedures, signal changes specific to the components of a functional task can be derived (see *Data preprocessing and extraction* below), while in the context of this study, we assume comparable hemodynamic response shapes of patients and controls.

fMRI data acquisition

For all three paradigms, imaging was performed with the same parameters using a 1.5 T Siemens Magnetom Avanto TIM-system MRI scanner (Siemens, Erlangen, Germany) equipped with a standard 12 channel head coil. In a single session, twenty-four 4-mm-thick, interleaved axial slices (in-plane resolution: 3.28 x 3.28 mm) oriented at the AC-PC transverse plane were acquired with 1 mm interslice gap, using a T2*-sensitive single-shot echo planar imaging (EPI) sequence with following parameters: repetition time (TR; 2000 ms), echo time (TE; 40 ms), flip angle (90°), matrix (64x64), and field of view (FOV; 210x210 mm^2). The first six volumes were discarded to account for magnetization saturation effects.

Data preprocessing and extraction

Data were preprocessed using the Statistical Parametric Mapping software (SPM5, Wellcome Department of Cognitive Neurology, UK). Slice-timing correction was applied, images were realigned, spatially normalized and smoothed, using an 8 mm FWHM Gaussian isotropic kernel. From this data, information for Gaussian process classification was extracted as follows:

For the first paradigm (viewing emotional faces), the mean value of the time series in each voxel was subtracted from each time point. Subsequently, the onset of each block was shifted by one TR to account for the hemodynamic delay and the average of eight consecutive volumes (i.e. the length of one block) was computed to construct a feature vector $\mathbf{x}_i$ (with dimensionality d equal to the number of voxels in the whole-brain mask). This procedure yielded four training samples – corresponding to the four repetitions of each block – for each facial expression. For each of these expressions, the test examples were created by averaging all training examples for each subject.

As the second and third paradigm (processing of reward and loss) were realized in a rapid event-related design, relevant information has to be extracted using a standard convolution model as implemented in SPM5 for each subject: Generally, a temporal overlap of the hemodynamic responses occurs if the stimuli are spaced closer together than the duration of the hemodynamic response cycle which returns to baseline after 10 to 12 seconds or more (Boynton, Engel, Glover, & Heeger, 1996; Buckner, et al., 1996). In order to obtain sufficient statistical power and avoid the induction of confounding states (e.g. boredom), however, a larger number of trials has to be presented in a randomized fashion while keeping total measurement time in a tolerable range. Thus, the question arises, how temporally overlapping hemodynamic responses following rapidly presented stimuli can be disentangled and

their amplitude quantified. Within the standard modeling approach applied in this work (Friston, et al., 1994; Friston, Jezzard, & Turner, 1994) it is assumed that stimulation will elicit neural firing followed by a hemodynamic response (*see Blood Oxygen Level Dependent (BOLD) imaging*). The shape of this hemodynamic response function (HRF) can be modeled using a Gaussian normal distribution. Provided successive hemodynamic responses summate linearly – as has been shown for inter-stimulus intervals of as low as 2 seconds (Dale & Buckner, 1998) – the General Linear Model (GLM) can be applied. In the GLM,[15] the fMRI time series of each voxel (data vector y) is predicted on the basis of a set of reasonable HRFs which are convolved with the event sequence of the experimental trials yielding the design matrix $\mathbf{x}$. To bind the variance induced by potential movement artifacts, measurements of movement in the 3 directions of translation and 3 degrees of rotation were added as regressors to the design matrix. Then data was modeled as $y = \mathbf{x}\beta + \varepsilon$, where β is the weight vector and ε represents an error term. Within this framework, the estimation of the β-weights quantifies the contribution of the HRF to the explanation of the data y which in the current context corresponds to an estimate of the amplitude of the hemodynamic response. Assuming ε to be uncorrelated, independent and normally distributed, the unbiased ordinary least square estimates of β are given by $\beta = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T y$. As, however, several physiological processes, such as respiration or blood-pressure changes, might pose confounding factors, the time series in each voxel were high pass filtered to $\frac{1}{128}$ Hz and corrected for temporal auto-correlation using an autoregressive model with a lag of 1 (Orcutt & Cochrane,

[15] It is important to note that common notation within the GLM differs from the notation regularly used in pattern recognition (used in Part I of this work). In contrast to pattern recognition, y denotes the single voxel time-series data (not the labels) while $\mathbf{x}$ represents the design matrix (not whole-brain data) mainly derived from the experimental setup. While a complete integration of the fMRI GLM model into GPR is possible (Marquand, et al., 2009), we use the common, though overlapping notation for both methods to avoid confusion.

1949) before β-weights were estimated. Following the rational based on the mechanisms of neurovascular coupling, β-weights might now be interpreted as brain activation (Friston, et al., 1994; Friston, et al., 1994).

In the rapid event-related framework of paradigms two and three (processing of reward and loss), the β-weights for the cue and feedback in each condition were estimated for each subject as described above. Specifically, this procedure yielded six β-maps per subject corresponding to the anticipation of large, small, and no reward or loss, respectively. Likewise, twelve β-maps corresponding to large, small, and no reward/loss feedback following a correct or an incorrect response, respectively, were obtained. If the number of trials on which a single subject's β-estimates were based was below six – due to e.g. an extremely low number of incorrect responses for a particular subject – this condition was excluded from further analysis for all subjects. Applying this criterion, valid data associated with actual small and large reward feedback as well as with avoidance of small and large losses could be obtained from the feedback phases of the second and third paradigm. By collapsing the data from actual small and large loss feedback, these two conditions could additionally be included. The feature vectors $\mathbf{x}_i$ (for each subject) were constructed directly from each condition's β-map.

To address the concern that classification results might be driven by differences in movement between patients and controls during the measurements, we used a GPC to classify patients and controls based on their movement parameters acquired during each paradigm. LOO-CV did not yield significant accuracies for any of the three paradigms.

Finally, for all three paradigms, a mask containing all intracerebral voxels was applied, yielding whole brain data for Gaussian process classification. In summary,

for each of the 60 subjects, data from 15 conditions (see Figure 3 for a complete list) could be extracted for use in classification.

3.3.4. Algorithm application

Whole-brain data from the 15 conditions and 60 subjects were analyzed using the multi-source pattern classification algorithm outlined in this work (see 2.3 Algorithm development). Specifically, the feature vector $\mathbf{x}_i$ for each subject consisted of all (d) intracerebral voxels measured in the respective condition (see *Data preprocessing and extraction* for details). In order to benchmark the single GP classifiers' performances, we compared GP classifier accuracies to the performance of linear support vector machine classifiers which constitute the most widely used pattern recognition approach in the field of neuroimaging (Fu, et al., 2008; Marquand, et al., 2008; Mourao-Miranda, et al., 2005).

In terms of coding and implementation, GP classification was performed using a customized version of the Gaussian processes for machine learning (GPML) toolbox for Matlab (http://www.gaussianprocess.org/gpml). We used a linear covariance function and estimated hyper-parameters controlling bias and regularization using an empirical Bayesian approach (for details, see 2.3 Algorithm development). To improve visualization of brain regions important for classification and eliminate noise components, we thresholded all multivariate discrimination maps at 30% of the maximum intensity and 5 continuous voxels (for successful application of weight-map thresholding, see e.g. Costafreda, et al., 2009). SVM calculations were done using the LIBSVM library for Matlab (http://www.csie.ntu.edu.tw/~cjlin/libsvm) where classification was performed using a linear kernel. In all cases, we fixed the SVM

parameter C which controls the trade-off between maximizing the margin and permitting misclassification to the default value of one.

Particularly in the context of neuroimaging, it is of great interest which regions contribute most to classification. As described in detail above (2.3.4 Multivariate feature mapping), we can differentiate between two general mapping approaches: one which quantifies the contribution of each voxel to the decision boundary (weight maps) and one showing the distribution of the two groups relative to one-another with respect to this decision boundary (distribution maps). Thus, weight maps derived from neuroimaging data are a spatial representation of a GP classifier's decision boundary while group maps constitute a spatial quantification of the difference between the two classes, i.e. subjects classified as patients or controls. Based on this, node-specific spatial distribution maps, as reported here (3.4.3 Multivariate spatial mapping of neural processes), can be understood to show those brain regions on which subjects classified at each node of the decision tree differ most relative to the decision boundary (a list containing information on all regions contributing to the formation of the decision boundary for each relevant biomarker can be found in the Appendix in Table A - 1, Table A - 2, and Table A - 3; for computational details, see 2.3.4 Multivariate feature mapping).

3.4. Results

3.4.1. Classification based on single biomarkers

Independent GP classification of the data from each of the 15 conditions revealed significant accuracies for a total of 8 conditions (Figure 3).[16] The median accuracy for

[16] As we did not intend to draw inferences based on the performance of any single classifier, we did not perform multiple comparison correction for the 15 tests.

all GP classifiers was at 60% while the single best classifier (*anticipation of no loss*) performed at an accuracy of 72%.

Furthermore, we compared the GP classifiers to the standard SVM approach. Generally, both algorithms performed comparably, with slight advantages for the GP classifiers which reached accuracies at least as high as the SVM in all but one of the 15 cases (for an overview, see Figure 3).

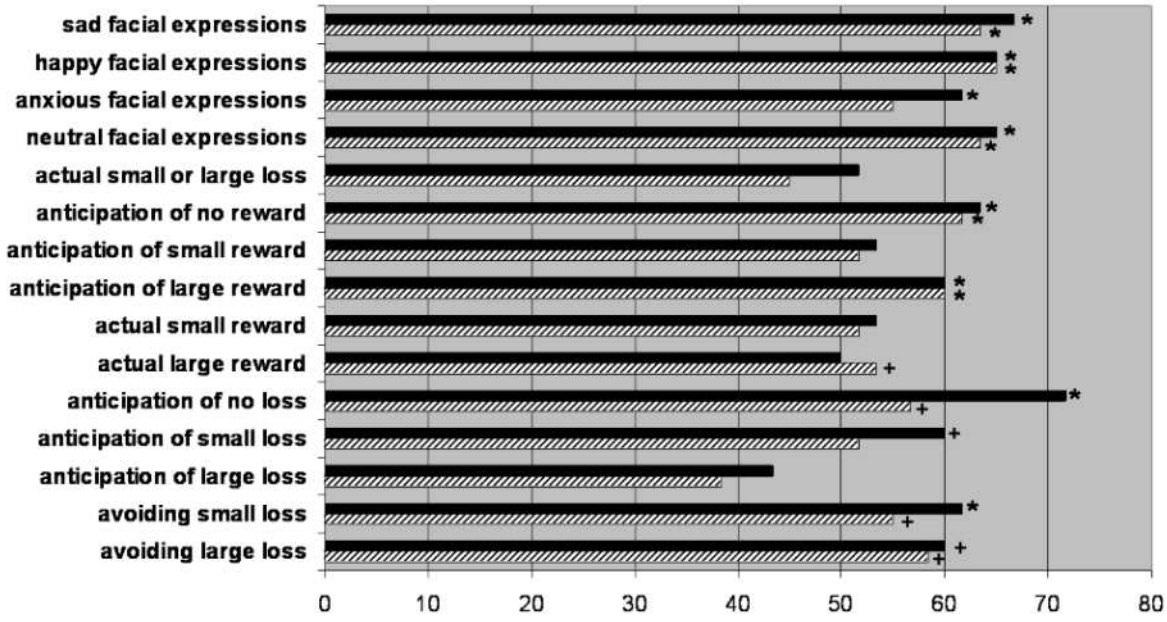

Figure 3. Gaussian Process classification accuracies in percent (black bars) for all 15 conditions compared to support-vector machine accuracies (shaded bars). * and + indicate above chance-level classification accuracy (*: $p < .05$; +: $p < .10$).

3.4.2. Integrated biomarker classification

Integrating the descriptive probabilities from all single GP classifiers using the decision tree algorithm leads to an accuracy of 83%. This constitutes an improvement in accuracy of 11% ($p=.017$) in comparison with the single best of all GP classifiers. The boost in accuracy compared to the median of all GP classifiers amounts to 23% (Figure 4).

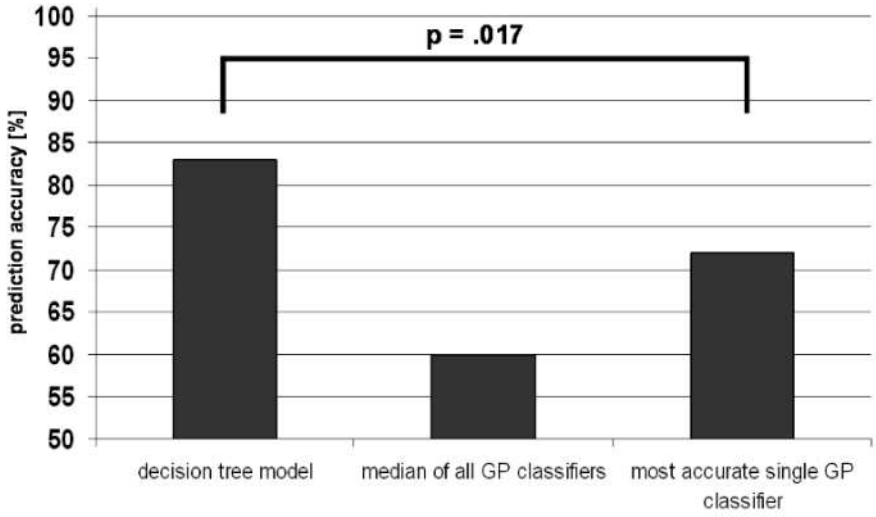

Figure 4. Increase in accuracy for the decision tree model integrating conditions (83%) in comparison to the median of all Gaussian Process (GP) classifiers (60%) and the most accurate single GP classifier alone (72%).

Investigating the optimal decision tree model revealed which conditions were relevant for overall classification (Figure 5): The entire subject group was best classified by splitting the p_{GP} for *neutral facial expressions* at .46. In the second step, those subjects with a p_{GP} for *neutral facial expressions* lower than .46 (left branch) were best classified by splitting the p_{GP} for *actual large reward* at .39. For those subjects more likely to be patients based on the p_{GP} for *neutral facial expressions* (right branch), the best classification was obtained based on the p_{GP} for *anticipation of no loss* splitting at .47.

In summary, integrating p_{GP} using a decision tree algorithm substantially boosted classification accuracy by considering GP predictive probabilities derived from three conditions. Note that these conditions are not those with the highest single GP classification accuracies. While there are three conditions related to the processing of emotional facial expressions among the four most accurate single GP classifiers, only the p_{GP} for *neutral facial expressions* is relevant for the construction of the tree

model. Furthermore, while the single GP classifier based on *actual large reward* does not classify the entire sample above chance level (see Figure 3), it nonetheless contains information essential for the classification of participants into subsamples.

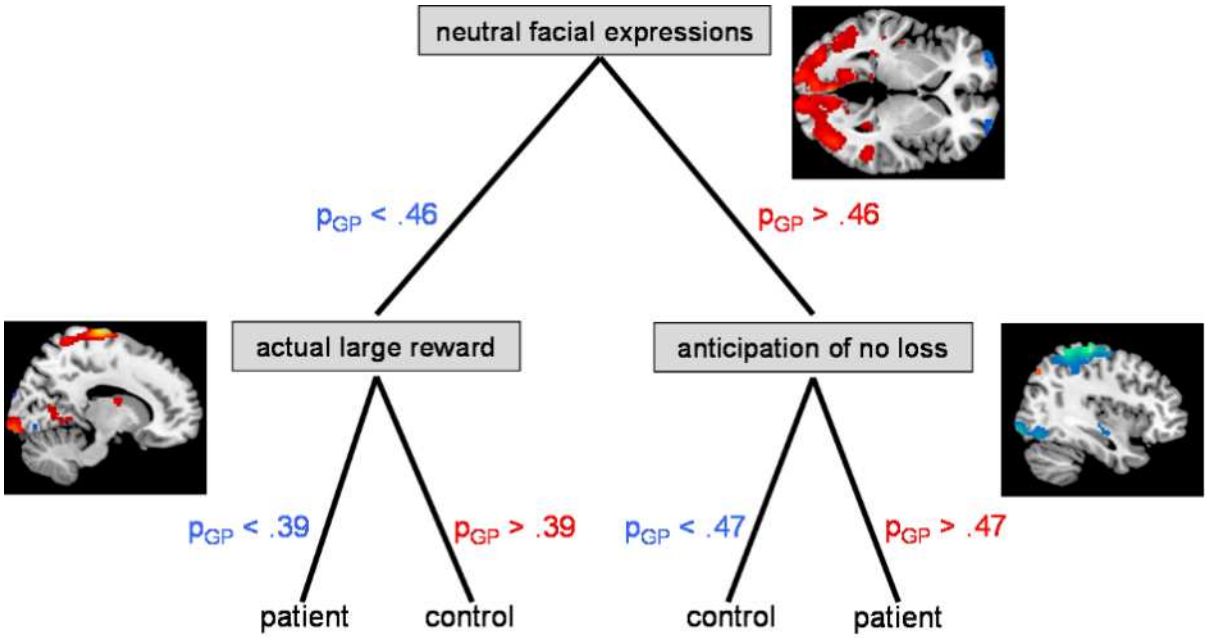

Figure 5. Optimal decision tree model showing variables relevant for overall classification. Subjects' predicted probabilities to be a patient (p_{GP}) derived from fMRI data related to the processing of neutral facial expressions was most informative for classification of the whole sample. Subjects in the two resulting subsamples could be classified best using p_{GP} derived from data related to reward (*actual large reward*) and safety (*anticipation of no loss*). Brain maps show node-specific distribution maps: shades of blue indicate classification to the left branch; shades of red indicate classification to the right branch.

3.4.3. Multivariate spatial mapping of neural processes

For all three biomarkers relevant for final prediction, a complex, intercorrelated pattern of regions was found to contribute to classification (Figure 5; node-specific maps; a complete list with information for all regions can be found in the Appendix in Table A - 4, Table A - 5, and Table A - 6). In the first classification step (split on

neutral facial expressions) this network includes a large occipital-parietal cluster containing the fusiform gyrus which contributes to classification to the right branch of the tree. In addition, smaller clusters within the caudate as well as in frontal regions show highest coefficient scores, suggestive of an important difference in these regions. While regions indicating classification to the left branch also include occipital regions (inferior occipital gyrus) and the lingual gyrus, mainly frontal regions (superior frontal and middle frontal gyrus) are essential.

Following the tree to the node-specific map for *anticipation of no loss*, we again find an extended occipital-parietal cluster comprising the cuneus, the lingual gyrus and the precentral gyrus. This time, however, it is characteristic of classification to the left branch (controls; as opposed to the right branch for *neutral facial expressions*). Likewise, the lingual gyrus now shows coefficient scores characteristic of classification to the right branch (patients).

Investigating the split on *actual large reward*, an extensive parietal cluster including the post central gyrus is characteristic of right-branch classification (controls) in this subsample. Furthermore, we find smaller clusters in superior temporal regions as well as the thalamus and the amygdala. Areas characteristic of left-branch classification (patients) are the cuneus and again the lingual gyrus.

4. Discussion

In this work, we sought to improve the predictive power of biomarkers of depression by combining multiple neurobiological markers. First, we identified the core-symptoms of depression from standard classification systems (3.1.1.2 Symptoms and diagnosis of depression). Then, we designed and conducted three experimental paradigms probing psychological processes known to be related to these symptoms (3.1.2 Biological markers of depression). In order to integrate the resulting 12 high-dimensional biomarkers (3.3.3 Functional Magnetic Resonance Imaging), we developed a multi-source pattern recognition algorithm based on a combination of GPC and CART (Part I – Integrating biomarkers: development of a multi-source pattern classification algorithm). Applying this method to a group of 30 healthy controls and 30 depressive in-patients who were on a variety of medications and displayed varying degrees of symptom-severity allowed for high-accuracy single-subject classification. Specifically, integrating biomarkers yielded an accuracy of 83% while the best of the 12 single biomarkers alone classified a significantly lower number of subjects (72%) correctly. Thus, integrated multi-source biomarker-based classification of a heterogeneous, real-life sample resulted in accuracy comparable to the highest ever achieved in previous single biomarker research (Fu, et al., 2008). In addition, the high-accuracy results reported by Fu et al. (2008) were obtained in specifically selected, homogeneous group of unmedicated participants.

Investigation of the final prediction model revealed that neural activation during the processing of neutral facial expressions, large rewards, and safety cues is most relevant for over-all classification. While *neutral facial expressions* were most informative, *actual large reward* and *anticipation of no loss* classified the resulting two subsamples optimally. This underlines the general strength of the algorithm to

specifically include non-redundant sources of information from a larger number of biomarkers into the model. We conclude that combining brain activation related to the core-symptoms of depression using the multi-source pattern classification approach developed in this work substantially increases classification accuracy while providing a sparse relational biomarker-model for future prediction.

4.1. Single biomarkers of depression

Investigating the single biomarkers acquired in this work, we were able to provide evidence showing that neural correlates of emotional processing – which had previously been identified as biomarkers of depression on the group-level – are also useful for single-subject classification. A total of eight single biomarkers classified the sample above chance level: half of these are based on neural responses to emotional facial expressions (*happy*, *sad*, *anxious*, and *neutral facial expressions*). The remaining four classifiers rely on data from *anticipation of no reward*, *anticipation of large reward*, *anticipation of no loss*, and *avoiding small loss*.

Neural responses during *anticipation of no loss* appear to have the highest predictive power. While behavioral studies specifically investigating the *anticipation of no loss* are missing, a neuroimaging study also using a version of the MID task showed activation in the middle and inferior frontal gyri as well as in parietal regions, the insula, caudate, and thalamus in depressive patients and controls. However, univariate analyses employed in this study did not show significant differences in these regions between depressive patients and healthy controls (Knutson, et al., 2008). Interestingly, the multivariate approach used in the current work reveals that all of the relevant regions identified by Knutson et al. (2008) – except for the thalamus – contribute to the formation of the decision boundary (compare decision

boundary weight map for *anticipation of no loss*, Appendix Table A - 3) and thus hold substantial predictive power. As the previous study did find those regions to play a role in the processing during the anticipation of no loss, we can only assume that mass univariate statistical testing with a sample size of 26 (14 depressive patients and 12 controls), as conducted in the previous study, provided too little statistical power to detect significant differences between patients and controls. Alternatively, it is conceivable that information differentiating the groups is coded in the interrelation of the regions which is disregarded by univariate approaches. From a theoretical point of view, the cue indicating no loss – independent of the subject's performance in this trial – might be interpreted as a sign of safety in a context of potential losses. As depressive patients have been shown to be more risk averse than healthy controls (Smoski, et al., 2008), such "safe trials" may appear more positive to depressive individuals than to controls. This also fits with findings by Elliott et al. (1998) concerning behavior in response to feedback who concluded "that depressive individuals generally behave as if they expected failure while healthy participants behave as if they expected success". When expecting to be unable to respond in time during the MID task (vs. expecting to succeed on that trial), a trial in which no loss can occur independent of the behavioral performance is much more positive. The same line of arguments might explain the significant predictive power based on *avoiding small loss*.

Considering the differences in reward-processing which have consistently been found between depressive individuals and controls (3.1.2 Biological markers of depression), the predictive power of neural responses during the anticipation of large and no rewards appears consistent with literature. Examining responses during the *anticipation of large reward*, we found highly similar regions to contribute to the formation of the decision boundary which have been identified in previous mass

univariate approaches as well: While previous studies identified the ventral medial prefrontal cortex, the amygdala, and the ventral striatum (Keedwell, Andrew, Williams, Brammer, & Phillips, 2005; Epstein, et al., 2006; Pizzagalli, et al., 2009), regions contributing to the formation of the decision boundary in our study included inferior, medial and superior frontal gyri, the amygdala as well as a large cluster within the caudate (Appendix Table A - 2). In analogy to those previous studies showing positive correlations of anhedonia with reward-related activation in frontal areas and negative correlations with the amygdala, we find that frontal areas contributed to the classification of patients while the amygdala classified controls.

Further investigation of single classifiers shows that neural responses to emotional faces provide consistently good classification accuracies for neutral, sad, anxious, and happy facial expressions. This is in line with the numerous findings concerning an attentional bias and altered behavioral responses to facial expressions such as higher error rates and longer reaction times for the identification of neutral, sad, and happy facial expressions (3.1.2.1 Processing of emotional stimuli; for reviews, see Leppanen, 2006 and Bylsma, et al., 2008). Considering previous studies using pattern recognition approaches, our results mirror findings related to sad and neutral facial expressions by Fu et al. (2008) while – at least nominally – showing lower accuracies. Anatomically, a highly similar pattern of discriminating regions for neutral faces is identified; including medial, middle and superior frontal gyri, the cuneus, the precuneus, lingual, precentral, postcentral, superior parietal as well as the parahippocampal gyri (Appendix Table A - 1). As the experimental conditions were very similar and as the decline in accuracy becomes even more evident if we use the same pattern recognition method employed by Fu et al. (SVM as used for benchmarking in this work; Fu, et al., 2008) while the contributing regions are largely the same, we conclude that a comparable decision boundary was

learned. However, it did not lead to equally good performance in our sample. Comparing the carefully chosen, unmedicated, acutely depressive subjects with our heterogeneous sample, this does not come as a surprise. It underlines, however, the importance of using subjects from a population for which accuracy estimation is relevant.

From this we conclude that – as hypothesized – single GP classifiers based on neural correlates of the processing of emotional facial expressions as well as those based on data from reward- and loss-processing can classify depressive individuals and controls which have not previously been seen by the algorithm with significant accuracy. Arguing for the validity of the approach, the regions contributing to classification include those which have previously been reported for mass univariate approaches while largely replicating those previously identified using multivariate pattern recognition.

4.2. Combining symptom-related biomarkers of depression

In psychiatry, the basic diagnostic process involves obtaining information on the symptoms a patient displays and combining this data for each symptom to yield a basis for valid diagnosis. This is mirrored in the standard classification systems such as DSM-IV or ICD-10 which outline detailed symptom-based criteria necessary for the diagnosis of psychiatric disorders. Striving to improve the accuracy of current biomarker-based classification, we adopted the idea of gathering symptom-related data and combining it for final classification. While the symptoms required for diagnosis by standard classification systems are mainly assessed based on patients' self-reports and clinical observation of behaviors, we sought to measure and combine multiple symptom-related neurobiological markers. Thus, we first identified

the core-symptoms of depression from standard classification systems to then obtain data on the neural processes known to be associated with these symptoms. In order to combine the resulting high-dimensional neurobiological markers, we developed a multi-source pattern classification algorithm based on a combination of GPC and CART (Part I – Integrating biomarkers: development of a multi-source pattern classification algorithm).

In accordance with our main hypothesis combining multiple biological markers of depression significantly improved classification accuracy. Specifically, integrating the predictive probabilities obtained from each GP classifier using a decision tree resulted in an accuracy of 83%. This constitutes an 11% increase in accuracy compared to the most accurate of the single GP classifiers alone and a 23% increase in comparison to the median of all single classifiers (Figure 4). As during the entire process of classification the algorithm was never provided with a subject's true class label, but only applied the classification rule learned from the other participants, it is reasonable to assume that classifying new subjects will lead to comparable accuracies. In summary, integrated biomarker-based classification of our heterogeneous, real-life sample of patients and controls resulted in accuracy comparable to the highest ever achieved in previous single biomarker research (Fu, et al., 2008) while no longer evaluating performance based on a homogeneous group of unmedicated participants.

Investigating the optimal tree model underlying final prediction identified three biomarkers mainly driving this substantial improvement of single-subject classification accuracy: Specifically, neuroimaging data related to the processing of neutral facial expressions was most informative for classification of the whole sample. This result is in line with evidence showing altered processing of neutral facial expressions (Leppanen, et al., 2004) in depression. Also, the results mirror findings by Fu et al.

(2008) who reported neutral facial expressions to have highest predictive power to classify depressive patients. Furthermore, even though data from sad, happy, and neutral faces all provide relatively high accuracies on the single classifier level, incorporating information from other emotional facial expressions is no longer of significant utility after splitting the sample based on data derived from neutral facial expressions. It appears that data from sad, happy, and neutral facial expressions provide similar information so that additionally incorporating data from sad or happy facial expressions does not substantially increase accuracies in any of the two subsamples resulting from the split based on data from neutral faces.

Examining the node-specific map for *neutral facial expression* (Appendix Table A - 4) reveals that the two resulting groups differ most in brain regions which have previously been shown to be relevant in depression: In line with Fu et al. (2008), we found a large occipital-parietal network including the fusiform gyrus to contribute to classification. As an area essential for the processing of stimulus features, the fusiform gyrus has also been shown to be differentially activated in response to emotional facial expressions in depressive individuals and controls using mass univariate testing (Surguladze, et al., 2005). Furthermore, the involvement of clusters within the caudate as well as in frontal regions is in line with previous evidence from task-related functional (Fu, et al., 2008; Keedwell, et al., 2005; Knutson, et al., 2008; Epstein, et al., 2006; Pizzagalli, et al., 2009) as well as structural and metabolic data (Krishnan, et al., 1992; Bremner, et al., 2002; Ito, et al., 1996; Kennedy, et al., 1997) in depression.

Subjects in the two subsamples resulting from the split on *neutral facial expressions* can be classified best using data related to reward (*actual large reward*) and safety (*anticipation of no loss*; Figure 5). These results fit well with previous studies showing reward- and loss-related deviations on the behavioral level in

depression (Elliott, et al., 1996; Henriques, et al., 1994). Furthermore, we can – in the context of tree classification – analyze the interrelations between the multiple biomarkers: While single classifiers based on *happy*, *sad*, and *anxious facial expressions* have much higher predictive power than the one based on *actual large reward*, it is nonetheless the latter variable which is selected within the decision tree. Obviously, facial expressions data discriminates well between depressive individuals and controls (Figure 3), however, the information obtained from each single classifier is largely redundant (i.e. similar individuals are classified correctly and incorrectly based on this data). However, subjects who were misclassified by *neutral facial expression* alone can be classified correctly based on *actual large reward*. In this context, it appears noteworthy that even though the single GP classifier based on *actual large reward* does not classify the entire sample above chance level, it nonetheless contains information essential for the classification of participants into subsamples. This underlines a general strength of the decision tree which subdivides the sample into a number of subsamples. Thereby, information that did not possess significant predictive power for the whole dataset can very well be of importance in a subsample (for details, see 4.3 Methodological considerations).

From the mathematical construction of the node-specific distribution maps (2.3.4 Multivariate feature mapping), it follows that only those features (brain regions) will contribute to classification at a given node which are not redundant with respect to the subjects classified before (see above). Thus, the node-specific distribution maps for *actual large reward* and *anticipation of no loss* (Appendix Table A - 5 and Table A - 6) show regions which add new information to classification rather than those regions most relevant for the whole group. Thus, for the second split only regions relevant for the correction of previous misclassifications are central. While these maps cannot readily be compared to maps derived from existing approaches (e.g. t-

or F-maps or SVM w-maps), it is most striking that regions previously (on the first split) indicating classification to one side of the tree contribute to classification to the other side of the tree on the next level. This way, information complemental to that used in the previous split is important to classification even though the pattern of relevant regions is similar. Prominent exceptions are the contribution of the amygdala and the thalamus for classification at the *actual large reward* node which could not be found in the previous map pattern.

When speculating on why so highly similar regional patterns – containing mainly occipital-parietal and frontal regions – emerge at each node in spite of the fundamentally different tasks, at least two explanations come to mind: For one, all tasks conducted in this work require the processing of visual stimuli. As it has been shown that neural responses in depressive patients differ in areas relevant for the processing of stimulus features (Surguladze, et al., 2005), the classifiers might have learned to differentiate patients and controls based on alterations in early visual processing or top-down regulation of such mechanisms. Differences in spatial and/or temporal dynamics in these areas between the three relevant tasks (*neutral facial expression*, *actual large reward*, and *anticipation of no loss*) might then provide sufficient unique information at each node to enable high-accuracy classification. Classification based on electrophysiological data – such as EEG – with its superior temporal resolution might help to shed light on this issue. The other explanation for the highly similar patterns in all three variables might come from metabolic alterations: Especially in the occipital and frontal regions in question, differences in GABA and glutamate levels between depressive patients and controls have been observed (Sanacora, et al., 1999; Sanacora, et al., 2004; Hasler, et al., 2007; see also 3.1.2.2 Neuroimaging markers). This might impact glutamatergic neurotransmission directly as well as indirectly via structural changes (loss of tissue;

Bhagwagar, et al., 2008). Considering the central role of astrocytes for both GABAergic neurotransmission (through the GABA precursor glutamine) and the BOLD response (through neurovascular coupling; Rossi, 2006), these metabolic alterations in depressive patients might impact measurements during all paradigms and conditions. Although beyond the scope of this work, it would be highly interesting to examine the interaction of (persistent) metabolic changes and (dynamic) task-related alterations in depression.

In the ongoing debate concerning the potential neurobiological foundations of diagnostic categories defined in standard classification systems, it has often been questioned whether particular disorders constitute neurobiologically meaningful entities which are unique for every disorder (for an overview, see Davis, Hanson, & Altevogt, 2008). In contrast, research has been highly successful in identifying neural correlates of specific symptoms. In this work, we thus chose to focus on correlates of a specific pattern of symptoms (lowered mood and anhedonia) rather than attempt to directly investigate a more abstractly defined, single disorder. Based on this, we assessed multiple symptom-related neural processes in patients sharing current or recent depressive symptoms who had been diagnosed with one of three distinct mood disorders (Recurrent depressive disorder, Depressive episodes, and Bipolar affective disorder). Our results show that accurate classification is possible in such a diagnostically heterogeneous group, suggesting shared neural mechanisms related to altered affective and motivational processing in all patients who display (or have recently displayed) severe depressive symptoms. Underlining the stability and real-life utility of the approach, such a high-accuracy classification can be obtained even if patients are differently medicated and vary greatly in regard to current severity of symptoms. In summary, combining neurobiological markers related to the core-symptoms of depression using the multi-source pattern classification approach

developed in this work substantially increased classification accuracy while providing a sparse relational biomarker-model for future prediction. This model identified neural correlates of the processing of neutral facial expressions as well as reward- and loss-related biomarkers to be most relevant for over-all classification. Arguing for the validity of the rule learned by the classification algorithm, the regions essentially contributing to classification are those which have previously been shown to differ between depressive patients and controls using multivariate as well as mass univariate neuroimaging methods.

4.3. Methodological considerations

When developing the algorithm used in this work, the main goal was the construction of a procedure which would allow for single-subject classification based on multiple high-dimensional biomarkers. We achieved this by first reducing dimensionality of the problem space to then apply a non-linear classifier to the resulting data.

On the first level, we trained and tested with GP classifiers to determine a subject's probability to be a patient for each single biomarker independently using a LOO-CV procedure. As GP classifiers suitable for neuroimaging data are a relatively new development (first publication by Marquand, et al., 2009), we benchmarked our first-level GP classifiers by comparing their performances to SVM, the most widely used pattern recognition method in neuroimaging. In accordance with the only other study using the two approaches (Marquand, et al., 2009), we found comparable accuracies for GP classifiers and SVM. As largest margin classifiers such as SVM can be shown to optimally classify the training data (Vapnik, 1995), it follows that the decision boundaries learned by the GP classifiers are also close to optimal. In

addition, comparable accuracies obtained with LOO-CV imply comparable generalization of the classifier functions. Together with the generally good accuracies, it can be concluded that linear SVM and GP classifiers are not prone to overfitting even in high-dimensional datasets. Importantly, the similar performance of GP classifiers and SVM speaks to the stability and suitability of the approximations (expectation propagation estimate of the maximum marginal likelihood) and transformations (probit likelihood constrain on regression to obtain probabilistic output) used in GPC (see 2.3.1 First-level prediction) as SVM results can be computed in closed form and thus do not rely on such methods.

When performing dimensionality reduction as is done on the first level, information potentially relevant for classification is lost. As our approach projects approximately 150,000 data points per biomarker and subject onto a single dimension (the probability to be a patient), the loss of information might be extreme. Results show, however, that information relevant for classification is very well preserved. In this context, it is particularly intriguing that the approach does not make any prior assumptions about which features might be relevant while reducing dimensionality to the lowest possible value of 1. From this point of view, our algorithm turns out to be similar to other feature selection methods with the characteristic that dimensionality reduction is maximal. The main improvement in comparison to other methods of feature selection lies in the fact that due to using LOO-CV on both levels, one does not need to find a selection criterion on the training data alone. Classification of subjects is done solely based on first-level test data predictions. This way, whole-brain information is considered and can be analyzed using the mapping procedure introduced while dimensionality and sample size (if reasonable to estimate accuracy) are no longer problematic.

In the next step, CART partitions second-level space thereby considering only information for the next split that is not redundant. It follows that information which did not possess significant predictive power for the whole dataset can very well be of great importance in a subsample (as is the case for *actual large reward*). In addition to this, class boundaries which are fixed at p=.5 on the first level can then be optimized for each subsample independently, thereby improving classification accuracy.

As computations involved in learning the hyperparameters of the covariance function on the first level and finding a function which non-linearly partitions second-level space – while mathematically principled – are far too complex to be transparent to humans, it is of great importance to be able to investigate the processes driving classification. Therefore, we chose a second-level classifier producing a simple series of if-then conditions (tree model) which can easily be understood. Also, the idea of selecting the variable which leads to the largest reduction of node impurity (2.3.2 Second-level prediction) is intuitive.

Understanding the processes which drive GP classification is much more complex, though. It is hardly possible to fully grasp the interaction of first-level multivariate patterns and second-level tree classification. To make this over-all classification more transparent, we thus developed a method which quantifies the contribution of single features to classification at each node (2.3.4 Multivariate feature mapping). The resulting node-specific distribution maps enable the identification of the most discriminative properties of a biomarker in the context of the second-level biomarker model. Generally, regions shown on multivariate maps are not independent of each other, but constitute a meaningful pattern only together. Thus, all areas depicted form a highly complex net of interdependencies which needs to be interpreted cautiously. However, the node-specific maps identify those features

(brain regions) uniquely contributing at each partitioning of second-level space. Thus, brain regions relevant at each node are highlighted making the classifier function more accessible and speculations about neural mechanisms underlying classification feasible.

4.4. Limitations

When performing classification in the context of depression – particularly when introducing a new algorithm to do so – a number of issues need to be addressed. In the following, we will first consider open methodological questions to then outline potential issues in the clinical context.

Methodological issues

Basically, classification is the prediction of class labels based on a rule learned from a training dataset. Thus far, we assumed that we know the true labels of every sample used to train the classifier. When, for instance, predicting an experimental condition from the data, this assumption is always true, because the researcher who designed the experiment knows exactly which condition was presented at which point during the experiment. In biomarker research, this is not always the case. For instance, a subject diagnosed with depression by the examiner might in fact suffer from another disorder. Thus, all accuracy estimates given in this work assume that the class labels provided by the examiner are true, i.e. all patients truly suffer from depression while all controls are free from psychiatric disorders. We addressed this problem by making the process of diagnosing a patient as reliable and transparent as possible. Specifically, we used two experienced examiners in conjunction with standardized psychometric tools. Also, we recruited only controls which passed a

standardized screening for psychiatric disorders and who reported no other relevant physical problems (for details see 3.3.1 Participants). While diagnosis as conducted for this study is the gold standard in psychiatry, we cannot quantify the extent to which mislabeled cases are nonetheless present in the sample. Methodologically, a number of procedures have been suggested to solve the so-called imperfect reference problem of classification (for a review of current approaches see Rutjes, Reitsma, Coomarasamy, Khan, & Bossuyt, 2007). Generally, procedures correct the imperfect reference or construct a (more accurate) reference from two or more available references. The first approach is feasible if the degree of imperfection (i.e. reliability and validity of the reference) is known. For the second method, procedures (tests) are combined weighting each test with a measure of its quality in order to obtain higher accuracy of the reference. In a way, our approach of combining the judgments of two experienced clinicians aided by standardized assessment tools is an example of this second method. While it ought to increase reliability and validity of our reference, again, we cannot quantify the error inherent in this procedure. A third idea involves a highly pragmatic approach: The results of the index test (in our case the classifier) are examined with regard to their practical meaning and consistency using available evidence or theoretical assumptions. Investigating the relevant features (brain regions) and comparing the results with evidence from previous studies is an example of the third approach. In this context, it appears positive that the regions essentially contributing to classification in this work are those which have previously been shown to differ between depressive patients and controls. A quantitative comparison of activation patterns obtained from multivariate and mass univariate method would be needed in this case. However, no such method is currently available. In the future, meta-analytic forms of pattern recognition might help with such problems. In summary, we have sound reasons to believe that our

reference standard is valid. However, exact quantification of potential errors is not possible. Thus, the accuracies obtained in this work must be seen as the lower bound of the true accuracy achievable.

In addition to the imperfect reference problem, the effects of preprocessing the data are unknown. As we did not intend to optimize accuracy for this specific dataset, but to provide a method suitable for a wide range of applications involving high-dimensional functional neuroimaging markers, we did not use customized preprocessing, but applied the two most common forms of data preparation in neuroimaging: averaging for blocked designs and hemodynamic modeling for rapid event-related data (3.3.3 Functional Magnetic Resonance Imaging). If research ever identifies biomarkers useful for practical application using the algorithm proposed in this work, a principled comparison of different preprocessing procedures ought to be conducted. This might enable customized preprocessing for those (final) biomarkers which are to be used in a clinical context, potentially further improving accuracy. For this work, we can only conclude that the most common preprocessing steps have led to accuracies equal to or higher than all currently available methods.

When performing classification in the context of this work, we refer to the data extracted from the 12 conditions of the three paradigms as potential biomarkers. While this is true, the data from the 12 conditions is by no means independent. This is also underlined by the selection of variables relevant for CART classification which shows a high degree of redundancy particularly within the datasets obtained during the processing of facial expressions (3.4.2 Integrated biomarker classification). Against this background, we cannot exclude the possibility that certain biomarkers only hold predictive power in the context of the other conditions or even paradigms. For example, neutral faces might elicit different neural activation patterns when presented amongst happy, sad, and anxious faces then in the context of other

neutral faces or for instance images of objects. Likewise, a cue signaling safety from monetary loss will probably be perceived differently depending on the amount of money that can be lost in other trials. These effects do not allow the conclusion that neutral facial expressions in conjunction with the relevant conditions from the MID task alone are in any way sufficient for classification. In other words, neither the neural responses measured nor the mathematical representation of the pattern discriminating patients and controls are independent of the experimental context. While this seems trivial, it will be an important field of investigation for biomarker research to not only discover markers, but to construct experiments which elicit suitable functional dynamic processes including trait as well as state markers.

Clinical considerations

In this work, we classified depressive patients and healthy controls with high accuracy by combining information from multiple biomarkers. Thus, we have shown that, based on the data provided, the algorithm can differentiate patients and controls. In diagnostics, however, the problem in most cases is not binary classification between individuals with a disorder and those without it, but the challenge is to assign the correct diagnosis from a variety of possibilities. If the algorithm is to be helpful in a diagnostic context, its specificity to depression needs to be demonstrated. Based on the data obtained in this work, it might be possible that the algorithm learns a rule which does not classify depressive patients, but generally discriminates between healthy individuals and mentally ill patients. While this would raise highly interesting questions about shared mechanisms of psychiatric disorders, the algorithm’s lack of specificity might render it powerless in differential diagnostics. As we carefully designed tasks related to the core-symptoms of depression and identified regions relevant in depression to contribute substantially to classification,

we have sound reason to believe that the algorithm is specific to depression. However, only further studies including other, similar pathologies such as anxiety disorders can shed light onto this issue.

Regarding specificity, a similar issue arises from the inclusion of patients suffering from bipolar disorder as done in this work (5 of the 30 patients): While unipolar and bipolar patients present with similar symptoms during depressive phases, evidence suggests that bipolar depression is characterized by more psychotic symptoms as well as increased psychomotor retardation (for a comprehensive summary, see Gotlib & Hammen, 2009). Likewise, neurobiological studies report large overlap between the neural substrate relevant in unipolar and bipolar depression while the neural correlates of specific processes differ considerably (for in-depth review and discussion, see Phillips & Vieta, 2007). Assigning the same label to neurobiologically potentially different subjects (i.e. patients) might pose a problem to classification as the rule learned from a sample mainly consisting of unipolar depressive patients (25 of 30 subjects) might not generalize well to bipolar patients. This problem can also not be mitigated by the fact that an unknown number of patients currently diagnosed with unipolar depression might display symptoms of mania in the future and thus ought to be diagnosed with bipolar disorder as well. Based on the very high accuracy achieved in this work, however, it appears unlikely that excluding the relevant patients would have substantially improved classification accuracy. More importantly, though, mapping the areas relevant for classification at each node did not yield one or more distorted or uninterpretable maps as would be expected when mixing samples with arbitrary class labels, but revealed the involvement of regions which have previously been reported in depression. While the low number of bipolar patients does not permit statistically meaningful inferences, we furthermore investigated how the five bipolar patients were classified within the optimal tree

model. Interestingly, 4 out the 5 patients were classified to the same terminal node created from the split based on *actual large reward*. It appears that the tree model recognizes bipolar patients as showing a different neural response pattern and classifies them into a separate node accordingly. As the low number of bipolar patients does not allow for an interpretation, only a study containing sufficiently many bipolar patients would be able to shed light onto this issue (for a recently initiated collaboration addressing this, see 4.5 Future directions). Summarizing, patients suffering from bipolar disorder are mainly classified correctly and thus do not lead to a substantial attenuation of prediction accuracy. Additionally, the classifier appears to account for potentially altered neural responses by adjusting tree structure.

In psychiatric research, it is often impossible to randomly assign subjects to the experimental conditions. Thus, all differences or discriminant patterns found in a study such as ours might be induced by factors associated with but not causally related to the class labels, i.e. the disorder. When classifying medication-free controls and patients all of which are on a variety of medications, it is conceivable that the classifier learned to differentiate between medicated and unmedicated subjects. While we cannot directly address this concern, the fact that the patients were on a variety of medications with different mechanisms of action makes it unlikely that the classifier could have derived a reasonable rule from drug-associated neural response patterns. Also arguing against a rule based on drug effects, regions relevant for classification are highly similar to those found by Fu et al. (2008) who measured unmedicated patients. Furthermore, evidence suggests that neural responses in a number of regions become more similar to the patterns in healthy controls following pharmacological intervention or psychotherapy (Joe, et al., 2006; Fu, et al., 2008). This would obviously impair classification rather than foster it.

Beside these problems related to finding specific biomarkers with sufficient predictive power, the most fundamental limitation of biomarker-based classification in depression lies in the nature of psychiatric disorders. Generally, a diagnostic procedure based solely on biomarkers can never be a substitute for thorough clinical examination and assessment of behavioral and cognitive symptoms. Additionally, psychiatric disorders have subjective components and individual aspects relevant for all aspects of treatment. Especially for psychotherapy establishing a relationship between patient and clinician is central. Furthermore, a patient's individual view of the disorder and the context in which symptoms developed and persist are essential for planning and conducting a successful therapeutic intervention. Thus, the classification approach proposed in this work – even if with all current issues solved and pitfalls circumvented – could never be a substitute for in-depth psychiatric examination. However, the results of this work show the potential of biomarkers to be a substantial diagnostic aid. The decision to which extent such a tool is helpful will – for the foreseeable future – have to be made for each individual case by an experienced clinician.

4.5. Future directions

While classification algorithms such as the one outlined in this work are by no means meant as a substitute for a thorough clinical examination and a proper diagnostic process (see above), particularly the capability of this approach to model the interrelation of multiple neurobiological markers could be of great utility. This is especially true when investigating symptom-related neural processes rather than aiming for mere classification accuracy. In this context, we showed that classification specifically relies on data derived from neural mechanisms associated with neutral

facial expressions as well as with reward- and loss-related processes. Incorporating other promising markers such as imipramine binding or MRS data measuring GABA and glutamine concentrations (3.1.2.2 Neuroimaging markers) might further improve prediction accuracy. As CART handles probabilities as well as any other type of data at all levels of measurement, such low-dimensional markers could be introduced directly into second-level analysis. Without using GPC, this would provide information on whether a single marker adds relevant information to over-all classification.

Moreover, the algorithm enables imaging genetics analyses without constrains on the number of genes or imaging paradigms as alpha inflation due to multiple comparisons and non-linear gene-gene or gene-image interactions are no longer problematic. If a smaller number of potential markers (e.g. certain single nucleotide polymorphisms) are of interest, first-level GPC can be omitted, directly entering the data into second-level CART. This procedure would additionally reveal interactions between genes and neuroimaging markers as well as between multiple genes. Such investigations based on the dataset used in this work are currently underway. The major limitation in this context would, however, be that no gene or brain region could be understood on its own, but only in the context of the multivariate interdependencies relevant for prediction.

It would, furthermore be particularly interesting to add psychometric markers and analyze which biomarkers become redundant. This might hint at the psychological processes associated with each marker.

Introducing biomarkers specific to different disorders, the algorithm ought to be able to differentiate between multiple disorders. Patients should be classified based on those biomarkers (i.e. at those nodes) relevant for their specific disorder. In this context, CART can be used as a cluster analytic method: classifying patients with different disorders ought to result in nodes (i.e. clusters) containing patients suffering

from one disorder. Such an approach would also be highly interesting regarding the ongoing debate concerning the potential neurobiological foundations of diagnostic categories defined in standard classification systems. As it has often been questioned whether particular disorders constitute neurobiologically meaningful entities (for an overview, see Davis, et al., 2008), such a method which clusters patients with different ICD-10 or DSM-IV diagnoses might help to identify groups of patients more homogeneous in regard to biological processes.

While a necessary step to identify potential biomarkers of depression and demonstrate the predictive power of the new algorithm, classifying depressive patients and healthy controls might not be the research question most valuable for future practical application. Predictions for disorders in which misclassification of patients is more frequent would be of much greater utility. In this context, we have recently initiated projects which aim to differentiate between subtypes of attention deficit hyperactivity disorders (ADHD) as well as between unipolar and bipolar patients.[17] While data of this work points toward possible high prediction accuracy for the differentiation between bipolar and unipolar depressive patients, only the now available larger sample can provide more conclusive evidence. In the long term, we furthermore aim to predict treatment response in anxiety disorders based on a combination of fMRI BOLD and arterial spin labeling (ASL) measurements.

Also, the algorithm is of further methodological interest: As our approach generally allows for the integration of multiple data sources, it can be used to combine data from different acquisition technologies with different sampling rates and dimensionalities such as EEG and functional Near-Infrared Spectroscopy (fNIRS), fNIRS and fMRI, as well as ASL and fMRI. When aiming for classification accuracy

[17] Projects will be conducted in collaboration with Dr. M. Schecklmann (Würzburg, ADHD) and Dr. M. Pyka (Marburg, uni-/bipolar classification). A Wellcome Trust project combining multiple inputs to classifiers has also been initiated (Prof. Shaw-Taylor, London).

for a specific purpose rather than the identification of interpretable processes, optimizing the input data and selecting the most discriminant features appears most promising.[18] To enable the classification of conditions rather than groups of subjects, which is particularly important for basic research, we have already implemented a suitable add-on to the current algorithm.

In summary, we developed an algorithm which is able to integrate multiple high-dimensional biological markers. Applying this method to classify depressive patients and healthy control subjects, we substantially improved the predictive power of biomarkers of depression. Furthermore, investigation of the final prediction model revealed that neural activation during the processing of neutral facial expressions, large rewards, and safety cues is most relevant for over-all classification. We conclude that combining brain activation related to the core-symptoms of depression using the multi-source pattern classification approach developed in this work substantially increases classification accuracy while providing a sparse relational biomarker-model for future prediction. In light of these findings, it appears that biomarker research in conjunction with recent methodological advancements has brought practical application of biomarkers as diagnostic aids within reach.

[18] The respective projects have been initiated in collaboration with Dr. A.-C. Ehlis (Würzburg; EEG and fNIRS), S. Heinzel (Würburg; fMRI and fNIRS), and in the framework of a project of the Interdisziplinäres Zentrum für klinische Forschung (IZKF)

5. References

Alpers, G. W., & Pauli, P. (2001). *Anxiety Sensitivity Index*: University of Würzburg.

APA (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington DC: American Psychiatric Association.

Bandelow, B. (1997). *Panik und Agoraphobie-Skala (PAS). Handanweisung*. Göttingen: Hogrefe.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *J Pers Assess, 67*(3), 588-597.

Bhagwagar, Z., Wylezinska, M., Jezzard, P., Evans, J., Ashworth, F., Sule, A., et al. (2007). Reduction in occipital cortex gamma-aminobutyric acid concentrations in medication-free recovered unipolar depressed and bipolar subjects. *Biol Psychiatry, 61*(6), 806-812.

Bhagwagar, Z., Wylezinska, M., Jezzard, P., Evans, J., Boorman, E., P, M. M., et al. (2008). Low GABA concentrations in occipital cortex and anterior cingulate cortex in medication-free, recovered depressed patients. *Int J Neuropsychopharmacol, 11*(2), 255-260.

Biomarkers and surrogate endpoints: preferred definitions and conceptual framework (2001). *Clin Pharmacol Ther, 69*(3), 89-95.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning* Berlin: Springer.

Blazer, D. G., Kessler, R. C., McGonagle, K. A., & Swartz, M. S. (1994). The prevalence and distribution of major depression in a national community sample: the National Comorbidity Survey. *Am J Psychiatry, 151*(7), 979-986.

Bode, S., & Haynes, J. D. (2009). Decoding sequential stages of task preparation in the human brain. *NeuroImage, 45*(2), 606-613.

Bostwick, J. M., & Pankratz, V. S. (2000). Affective disorders and suicide risk: a reexamination. *Am J Psychiatry, 157*(12), 1925-1932.

Botteron, K. N., Raichle, M. E., Drevets, W. C., Heath, A. C., & Todd, R. D. (2002). Volumetric reduction in left subgenual prefrontal cortex in early onset depression. *Biol Psychiatry, 51*(4), 342-344.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci, 16*(13), 4207-4221.

Breiman, L. (1996). Some Properties of Splitting Criteria. *Machine Learning, 24*, 41-47.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Bremner, J. D., Innis, R. B., Salomon, R. M., Staib, L. H., Ng, C. K., Miller, H. L., et al. (1997). Positron emission tomography measurement of cerebral metabolic correlates of tryptophan depletion-induced depressive relapse. *Arch Gen Psychiatry, 54*(4), 364-374.

Bremner, J. D., Narayan, M., Anderson, E. R., Staib, L. H., Miller, H. L., & Charney, D. S. (2000). Hippocampal volume reduction in major depression. *Am J Psychiatry, 157*(1), 115-118.

Bremner, J. D., Vythilingam, M., Vermetten, E., Nazeer, A., Adil, J., Khan, S., et al. (2002). Reduced volume of orbitofrontal cortex in major depression. *Biol Psychiatry, 51*(4), 273-279.

Bronisch, T., & Wittchen, H. U. (1994). Suicidal ideation and suicide attempts: comorbidity with depression, anxiety disorders, and substance abuse disorder. *Eur Arch Psychiatry Clin Neurosci, 244*(2), 93-98.

Brugha, T. S., Bebbington, P. E., MacCarthy, B., Sturt, E., Wykes, T., & Potter, J. (1990). Gender, social support and recovery from depressive disorders: a prospective clinical study. *Psychol Med, 20*(1), 147-156.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., et al. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc Natl Acad Sci U S A, 93*(25), 14878-14883.

Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clin Psychol Rev, 28*(4), 676-691.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). *An empirical evaluation of supervised learning in high dimensions*. Paper presented at the Proceedings of the 25th international conference on Machine learning.

Colombel, F. (2007). [Memory bias and depression: a critical commentary]. *Encephale, 33*(3 Pt 1), 242-248.

Costafreda, S. G., Chu, C., Ashburner, J., & Fu, C. H. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One, 4*(7), e6353.

Costafreda, S. G., Khanna, A., Mourao-Miranda, J., & Fu, C. H. (2009). Neural correlates of sad faces predict clinical remission to cognitive behavioural therapy in depression. *Neuroreport, 20*(7), 637-641.

Cross-national comparisons of the prevalences and correlates of mental disorders. WHO International Consortium in Psychiatric Epidemiology (2000). *Bull World Health Organ, 78*(4), 413-426.

Dai, H., Srikant, R., & Zhang, C. (2004). *Advances in knowledge discovery and data mining: 8th Pacific-Asia conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004 : proceedings*. Berlin ; New York: Springer.

Dale, A. M., & Buckner, R. L. (1998). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping, 5*(5), 11.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., et al. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *NeuroImage, 28*(3), 663-668.

Davatzikos, C., Shen, D., Gur, R. C., Wu, X., Liu, D., Fan, Y., et al. (2005). Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch Gen Psychiatry, 62*(11), 1218-1227.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Putnam, K. (2002). Depression: perspectives from affective neuroscience. *Annu Rev Psychol, 53*, 545-574.

Davis, M., Hanson, S., & Altevogt, B. (2008). *Neuroscience Biomarkers and Biosignatures: Converging Technologies, Emerging Partnerships*. Washington, D.C.: The National Academies Press.

Day, J. J., & Carelli, R. M. (2007). The nucleus accumbens and Pavlovian reward learning. *Neuroscientist, 13*(2), 148-159.

Delgado, P. L., Charney, D. S., Price, L. H., Aghajanian, G. K., Landis, H., & Heninger, G. R. (1990). Serotonin function and the mechanism of antidepressant action. Reversal of antidepressant-induced remission by rapid depletion of plasma tryptophan. *Arch Gen Psychiatry, 47*(5), 411-418.

Drevets, W. C. (2001). Neuroimaging and neuropathological studies of depression: implications for the cognitive-emotional features of mood disorders. *Curr Opin Neurobiol, 11*(2), 240-249.

Drevets, W. C., Frank, E., Price, J. C., Kupfer, D. J., Greer, P. J., & Mathis, C. (2000). Serotonin type-1A receptor imaging in depression. *Nucl Med Biol, 27*(5), 499-507.

Drevets, W. C., Price, J. L., Simpson, J. R., Jr., Todd, R. D., Reich, T., Vannier, M., et al. (1997). Subgenual prefrontal cortex abnormalities in mood disorders. *Nature, 386*(6627), 824-827.

Dunlop, B. W., & Nemeroff, C. B. (2007). The role of dopamine in the pathophysiology of depression. *Arch Gen Psychiatry, 64*(3), 327-337.

Elliott, R., Sahakian, B. J., McKay, A. P., Herrod, J. J., Robbins, T. W., & Paykel, E. S. (1996). Neuropsychological impairments in unipolar depression: the influence of perceived failure on subsequent performance. *Psychol Med, 26*(5), 975-989.

Elliott, R., Sahakian, B. J., Michael, A., Paykel, E. S., & Dolan, R. J. (1998). Abnormal neural response to feedback on planning and guessing tasks in patients with unipolar depression. *Psychol Med, 28*(3), 559-571.

Ellis, P. M., & Salmond, C. (1994). Is platelet imipramine binding reduced in depression? A meta-analysis. *Biol Psychiatry, 36*(5), 292-299.

Epstein, J., Pan, H., Kocsis, J. H., Yang, Y., Butler, T., Chusid, J., et al. (2006). Lack of ventral striatal response to positive stimuli in depressed versus normal subjects. *Am J Psychiatry, 163*(10), 1784-1790.

Evans, S. J., Choudary, P. V., Neal, C. R., Li, J. Z., Vawter, M. P., Tomita, H., et al. (2004). Dysregulation of the fibroblast growth factor system in major depression. *Proc Natl Acad Sci U S A, 101*(43), 15506-15511.

Filippi, M. (2009). *Fmri techniques and protocols*. New Jersey, NJ: Humana Press.

Fox, P. T., Raichle, M. E., Mintun, M. A., & Dence, C. (1988). Nonoxidative glucose consumption during focal physiologic neural activity. *Science, 241*(4864), 462-464.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*(4), 189-210.

Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping, 1*(2), 153-171.

Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., et al. (2008). Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry, 63*(7), 656-662.

Fu, C. H., Williams, S. C., Cleare, A. J., Scott, J., Mitterschiffthaler, M. T., Walsh, N. D., et al. (2008). Neural responses to sad facial expressions in major depression following cognitive behavioral therapy. *Biol Psychiatry, 64*(6), 505-512.

Gefvert, O., Lundberg, T., Wieselgren, I. M., Bergstrom, M., Langstrom, B., Wiesel, F., et al. (2001). D(2) and 5HT(2A) receptor occupancy of different doses of quetiapine in schizophrenia: a PET study. *Eur Neuropsychopharmacol, 11*(2), 105-110.

Gilboa-Schechtman, E., Presburger, G., Marom, S., & Hermesh, H. (2005). The effects of social anxiety and depression on the evaluation of facial crowds. *Behav Res Ther, 43*(4), 467-474.

Gotlib, I. H., & Hammen, C. L. (2009). *Handbook of depression* (2nd ed.). New York: Guilford Press.

Gotlib, I. H., Kasch, K. L., Traill, S., Joormann, J., Arnow, B. A., & Johnson, S. L. (2004). Coherence and specificity of information-processing biases in depression and social phobia. *J Abnorm Psychol, 113*(3), 386-398.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *J Abnorm Psychol, 113*(1), 121-135.

Gruenberg, A. M., Goldstein, R. D., & Pincus, H. A. (2005). Classification of Depression: Research and Diagnostic Criteria: DSM-IV and ICD-10. In M. L. Wong & J. Licinio (Eds.), *Biology of depression: From novel insights to therapeutic strategies*. Weinheim (Germany): Wiley-VCH.

Hahn, T., Dresler, T., Ehlis, A. C., Plichta, M. M., Heinzel, S., Polak, T., et al. (2009). Neural response to reward anticipation is modulated by Gray's impulsivity. *Neuroimage, 46*(4), 1148-1153.

Hamilton, M. (1960). A RATING SCALE FOR DEPRESSION. *Journal of Neurology, Neurosurgery & Psychiatry, 23*(1), 56-62.

Haro, J. M., Arbabzadeh-Bouchez, S., Brugha, T. S., de Girolamo, G., Guyer, M. E., Jin, R., et al. (2006). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *Int J Methods Psychiatr Res, 15*(4), 167-180.

Hasler, G., van der Veen, J. W., Tumonis, T., Meyers, N., Shen, J., & Drevets, W. C. (2007). Reduced prefrontal glutamate/glutamine and gamma-aminobutyric acid levels in major depression determined using proton magnetic resonance spectroscopy. *Arch Gen Psychiatry, 64*(2), 193-200.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *Das Beck Depressionsinventar II. Deutsche Bearbeitung und Handbuch zum BDI II.* Frankfurt a. M.: Harcourt Test Services.

Haynes, J. D. (2009). Decoding visual consciousness from human brain signals. *Trends Cogn Sci, 13*(5), 194-202.

Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nat Rev Neurosci, 3*(2), 142-151.

Hennings, J. M., Owashi, T., Binder, E. B., Horstmann, S., Menke, A., Kloiber, S., et al. (2009). Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - findings from the Munich Antidepressant Response Signature (MARS) project. *J Psychiatr Res, 43*(3), 215-229.

Henriques, J. B., & Davidson, R. J. (2000). Decreased responsiveness to reward in depression. *Emotion and Cognition, 14*(5), 13.

Henriques, J. B., Glowacki, J. M., & Davidson, R. J. (1994). Reward fails to alter response bias in depression. *J Abnorm Psychol, 103*(3), 460-466.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*(3), 199-236.

Huang, T. M., & Kecman, V. (2005). Gene extraction for cancer diagnosis by support vector machines--an improvement. *Artif Intell Med, 35*(1-2), 185-194.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park: Sage Publications.

Husain, M. M., McDonald, W. M., Doraiswamy, P. M., Figiel, G. S., Na, C., Escalona, P. R., et al. (1991). A magnetic resonance imaging study of putamen nuclei in major depression. *Psychiatry Res, 40*(2), 95-99.

Iadecola, C. (2004). Neurovascular regulation in the normal brain and in Alzheimer's disease. *Nat Rev Neurosci, 5*(5), 347-360.

Iosifescu, D. V., Greenwald, S., Devlin, P., Mischoulon, D., Denninger, J. W., Alpert, J. E., et al. (2009). Frontal EEG predictors of treatment outcome in major depressive disorder. *Eur Neuropsychopharmacol, 19*(11), 772-777.

Ito, H., Kawashima, R., Awata, S., Ono, S., Sato, K., Goto, R., et al. (1996). Hypoperfusion in the limbic system and prefrontal cortex in depression: SPECT with anatomic standardization technique. *J Nucl Med, 37*(3), 410-414.

Joe, A. Y., Tielmann, T., Bucerius, J., Reinhardt, M. J., Palmedo, H., Maier, W., et al. (2006). Response-dependent differences in regional cerebral blood flow changes with citalopram in treatment of major depression. *J Nucl Med, 47*(8), 1319-1325.

Karege, F., Perret, G., Bondolfi, G., Schwald, M., Bertschy, G., & Aubry, J. M. (2002). Decreased serum brain-derived neurotrophic factor levels in major depressed patients. *Psychiatry Res, 109*(2), 143-148.

Keedwell, P. A., Andrew, C., Williams, S. C., Brammer, M. J., & Phillips, M. L. (2005). The neural correlates of anhedonia in major depressive disorder. *Biol Psychiatry, 58*(11), 843-853.

Kendler, K. S., Walters, E. E., & Kessler, R. C. (1997). The prediction of length of major depressive episodes: results from an epidemiological sample of female twins. *Psychol Med, 27*(1), 107-117.

Kennedy, S. H., Javanmard, M., & Vaccarino, F. J. (1997). A review of functional neuroimaging in mood disorders: positron emission tomography and depression. *Can J Psychiatry, 42*(5), 467-475.

Kessler, R. C. (1997). The effects of stressful life events on depression. *Annu Rev Psychol, 48*, 191-214.

Kessler, R. C., Akiskal, H. S., Ames, M., Birnbaum, H., Greenberg, P., Hirschfeld, R. M., et al. (2006). Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers. *Am J Psychiatry, 163*(9), 1561-1568.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA, 289*(23), 3095-3105.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry, 62*(6), 593-602.

Kessler, R. C., & Merikangas, K. R. (2004). The National Comorbidity Survey Replication (NCS-R): background and aims. *Int J Methods Psychiatr Res, 13*(2), 60-68.

Kessler, R. C., Soukup, J., Davis, R. B., Foster, D. F., Wilkey, S. A., Van Rompay, M. M., et al. (2001). The use of complementary and alternative therapies to treat anxiety and depression in the United States. *Am J Psychiatry, 158*(2), 289-294.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci, 21*(16), RC159.

Knutson, B., Bhanji, J. P., Cooney, R. E., Atlas, L. Y., & Gotlib, I. H. (2008). Neural responses to monetary incentives in major depression. *Biol Psychiatry, 63*(7), 686-692.

Krishnan, K. R., McDonald, W. M., Escalona, P. R., Doraiswamy, P. M., Na, C., Husain, M. M., et al. (1992). Magnetic resonance imaging of the caudate

nuclei in depression. Preliminary observations. *Arch Gen Psychiatry, 49*(7), 553-557.

Krohne, H. W., Egloff, B., Kohlmann, C. W., & Tausch, A. (1996). Untersuchungen mit einer Deutschen Version der “Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*, 139-156.

Kuss, M., & Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research, 6*, 1679-1704.

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das State-Trait-Angstinventar (STAI)*. Weinheim: Beltz.

Leppanen, J. M. (2006). Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. *Curr Opin Psychiatry, 19*(1), 34-39.

Leppanen, J. M., Milders, M., Bell, J. S., Terriere, E., & Hietanen, J. K. (2004). Depression biases the recognition of emotionally neutral faces. *Psychiatry Res, 128*(2), 123-133.

Lopez, A. D., & Murray, C. C. (1998). The global burden of disease, 1990-2020. *Nat Med, 4*(11), 1241-1243.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience.

MacQueen, G. M. (2009). Magnetic resonance imaging and prediction of outcome in patients with major depressive disorder. *J Psychiatry Neurosci, 34*(5), 343-349.

Maes, M., Meltzer, H. Y., Bosmans, E., Bergmans, R., Vandoolaeghe, E., Ranjan, R., et al. (1995). Increased plasma concentrations of interleukin-6, soluble interleukin-6, soluble interleukin-2 and transferrin receptor in major depression. *J Affect Disord, 34*(4), 301-309.

Maes, M., Meltzer, H. Y., Buckley, P., & Bosmans, E. (1995). Plasma-soluble interleukin-2 and transferrin receptor in schizophrenia and major depression. *Eur Arch Psychiatry Clin Neurosci, 244*(6), 325-329.

Manji, H. K., Drevets, W. C., & Charney, D. S. (2001). The cellular neurobiology of depression. *Nat Med, 7*(5), 541-547.

Marquand, A. F., Howard, M., Brammer, M., Chu, C., Coen, S., & Mourao-Miranda, J. (2009). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage*.

Marquand, A. F., Mourao-Miranda, J., Brammer, M. J., Cleare, A. J., & Fu, C. H. (2008). Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. *Neuroreport, 19*(15), 1507-1511.

Mayberg, H. S., Brannan, S. K., Mahurin, R. K., Jerabek, P. A., Brickman, J. S., Tekell, J. L., et al. (1997). Cingulate function in depression: a potential predictor of treatment response. *Neuroreport, 8*(4), 1057-1061.

Mellerup, E. T., & Plenge, P. (1988). Imipramine binding in depression and other psychiatric conditions. *Acta Psychiatr Scand Suppl, 345*, 61-68.

Meltzer, H. Y., & Arora, R. C. (1986). Platelet markers of suicidality. *Ann N Y Acad Sci, 487*, 271-280.

Milak, M. S., Parsey, R. V., Keilp, J., Oquendo, M. A., Malone, K. M., & Mann, J. J. (2005). Neuroanatomic correlates of psychopathologic components of major depressive disorder. *Arch Gen Psychiatry, 62*(4), 397-408.

Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet, 374*(9690), 609-619.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *Br J Psychiatry, 134*, 382-389.

Mössner, R., Mikova, O., Koutsilieri, E., Saoud, M., Ehlis, A. C., Muller, N., et al. (2007). Consensus paper of the WFSBP Task Force on Biological Markers: biological markers in depression. *World J Biol Psychiatry, 8*(3), 141-174.

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage, 28*(4), 980-995.

Murray, C. J., & Lopez, A. D. (1996). Evidence-based health policy--lessons from the Global Burden of Disease Study. *Science, 274*(5288), 740-743.

Nestler, E. J., & Carlezon, W. A., Jr. (2006). The mesolimbic dopamine reward circuit in depression. *Biol Psychiatry, 59*(12), 1151-1159.

Nickisch, H., & Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research, 9*, 2035-2078.

Orcutt, G. H., & Cochrane, D. (1949). A sampling study of the merits of autoregressive and reduced form transformation in regression analysis. *J Am Stat Assoc, 44*(247), 356-372.

Oswald, D. W., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT)* (2 ed.). Göttingen: Hogrefe.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage, 45*(1 Suppl), S199-209.

Pezawas, L., Meyer-Lindenberg, A., Drabant, E. M., Verchinski, B. A., Munoz, K. E., Kolachana, B. S., et al. (2005). 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *Nat Neurosci, 8*(6), 828-834.

Phillips, M. L., & Vieta, E. (2007). Identifying functional neuroimaging biomarkers of bipolar disorder: toward DSM-V. *Schizophr Bull, 33*(4), 893-904.

Pizzagalli, D. A., Holmes, A. J., Dillon, D. G., Goetz, E. L., Birk, J. L., Bogdan, R., et al. (2009). Reduced caudate and nucleus accumbens response to rewards in unmedicated individuals with major depressive disorder. *Am J Psychiatry, 166*(6), 702-710.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT press.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Education Psychology, 74*, 166-169.

Rossi, D. J. (2006). Another BOLD role for astrocytes: coupling blood flow to neural activity. *Nat Neurosci, 9*(2), 159-161.

Rottenberg, J., & Johnson, S. L. (2007). *Emotion and psychopathology : bridging affective and clinical science* (1st ed.). Washington, DC: American Psychological Association.

Rubinow, D. R., & Post, R. M. (1992). Impaired recognition of affect in facial expression in depressed patients. *Biol Psychiatry, 31*(9), 947-953.

Rutjes, A. W., Reitsma, J. B., Coomarasamy, A., Khan, K. S., & Bossuyt, P. M. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess, 11*(50), iii, ix-51.

Sanacora, G., Gueorguieva, R., Epperson, C. N., Wu, Y. T., Appel, M., Rothman, D. L., et al. (2004). Subtype-specific alterations of gamma-aminobutyric acid and glutamate in patients with major depression. *Arch Gen Psychiatry, 61*(7), 705-713.

Sanacora, G., Mason, G. F., Rothman, D. L., Behar, K. L., Hyder, F., Petroff, O. A., et al. (1999). Reduced cortical gamma-aminobutyric acid levels in depressed

patients determined by proton magnetic resonance spectroscopy. *Arch Gen Psychiatry, 56*(11), 1043-1047.

Schwenkmezger, P., Hodapp, V., & Spielberger, C. D. (1992). *Das State-Trait-Ärgerausdrucks-Inventar STAXI*. Bern: Huber.

Seminowicz, D. A., Mayberg, H. S., McIntosh, A. R., Goldapple, K., Kennedy, S., Segal, Z., et al. (2004). Limbic-frontal circuitry in major depression: a path modeling metanalysis. *NeuroImage, 22*(1), 409-418.

Sheline, Y. I., Gado, M. H., & Price, J. L. (1998). Amygdala core nuclei volumes are decreased in recurrent major depression. *Neuroreport, 9*(9), 2023-2028.

Sheline, Y. I., Sanghavi, M., Mintun, M. A., & Gado, M. H. (1999). Depression duration but not age predicts hippocampal volume loss in medically healthy women with recurrent major depression. *J Neurosci, 19*(12), 5034-5043.

Siegle, G. J., Carter, C. S., & Thase, M. E. (2006). Use of FMRI to predict recovery from unipolar depression with cognitive behavior therapy. *Am J Psychiatry, 163*(4), 735-738.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biol Psychiatry, 51*(9), 693-707.

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature, 460*(7252), 202-207.

Sluzewska, A., Rybakowski, J., Bosmans, E., Sobieska, M., Berghmans, R., Maes, M., et al. (1996). Indicators of immune activation in major depression. *Psychiatry Res, 64*(3), 161-167.

Smoski, M. J., Lynch, T. R., Rosenthal, M. Z., Cheavens, J. S., Chapman, A. L., & Krishnan, R. R. (2008). Decision-making and risk aversion among depressive adults. *J Behav Ther Exp Psychiatry, 39*(4), 567-576.

Surguladze, S., Brammer, M. J., Keedwell, P., Giampietro, V., Young, A. W., Travis, M. J., et al. (2005). A differential pattern of neural response toward sad versus happy facial expressions in major depressive disorder. *Biol Psychiatry, 57*(3), 201-209.

Surguladze, S. A., Young, A. W., Senior, C., Brebion, G., Travis, M. J., & Phillips, M. L. (2004). Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. *Neuropsychology, 18*(2), 212-218.

Suslow, T., Dannlowski, U., Lalee-Mentzel, J., Donges, U. S., Arolt, V., & Kersting, A. (2004). Spatial processing of facial emotion in patients with unipolar depression: a longitudinal study. *J Affect Disord, 83*(1), 59-63.

Torrubia, R., Ávila, C., Moltó, J., & Caseras, X. (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. *Personality and Individual Differences, 31*(6), 837-862.

Tremblay, L. K., Naranjo, C. A., Graham, S. J., Herrmann, N., Mayberg, H. S., Hevenor, S., et al. (2005). Functional neuroanatomical substrates of altered reward processing in major depressive disorder revealed by a dopaminergic probe. *Arch Gen Psychiatry, 62*(11), 1228-1236.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Wade, T. J., & Cairney, J. (2000). Major depressive disorder and marital transition among mothers: results from a national panel study. *J Nerv Ment Dis, 188*(11), 741-750.

WHO (1992). *The ICD-10 classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines* Geneva: World Health Organization

Wittchen, H.-U., Zaudig, M., & Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV*. Weinheim: Beltz.

Wittchen, H. U., & Jacobi, F. (2005). Size and burden of mental disorders in Europe--a critical review and appraisal of 27 studies. *Eur Neuropsychopharmacol, 15*(4), 357-376.

Zeng, T., & Liu, J. (2009). Mixture classification model based on clinical markers for breast cancer prognosis. *Artif Intell Med*.

6. Appendix

Table A - 1. Decision boundary weight map: ***neutral facial expressions***

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Fusiform Gyrus		-12.75	1304	[-46 -62 -20]
Lingual Gyrus		12.35	1208	[8 -98 -14]
Middle Occipital Gyrus	18	-9.99	1752	[44 -80 -16]
Superior Frontal Gyrus		-9.69	1560	[-18 66 10]
Superior Temporal Gyrus	22	9.55	3216	[66 -24 0]
Fusiform Gyrus	19	9.54	1080	[54 -66 -18]
Optic Tract		-9.48	576	[-6 0 -16]
Cuneus		-9.43	312	[-10 -104 -4]
Superior Temporal Gyrus		9.23	616	[60 4 -4]
Middle Occipital Gyrus		-8.97	3192	[30 -94 6]
Superior Frontal Gyrus	6	-8.89	504	[-4 16 68]
Precuneus	7	-8.57	2552	[4 -74 44]
Calcarine_R		-8.51	2000	[2 -76 6]
Postcentral Gyrus	2	8.50	440	[20 -34 76]
Cuneus		8.34	1376	[2 -86 34]
Inferior Parietal Lobule		-8.20	1536	[42 -50 58]
Cuneus	19	8.01	728	[-4 -88 36]
Superior Parietal Lobule	7	-7.86	424	[-26 -70 56]
Superior Frontal Gyrus	10	7.72	1504	[18 68 18]
Uncus	28	7.72	1232	[22 8 -32]
Superior Parietal Lobule	7	7.55	1784	[-36 -62 56]
Inferior Frontal Gyrus	47	7.54	2128	[-52 18 -6]
Frontal Lobe		7.14	680	[-18 32 -16]
Precentral Gyrus	6	7.14	408	[-8 -20 78]
Temporal_Mid_L		-7.13	1480	[-46 -2 -20]
Frontal_Sup_L		-7.00	776	[-20 -4 74]
Fusiform Gyrus	37	6.99	608	[50 -44 -20]
Middle Frontal Gyrus		6.98	2816	[-40 50 -10]
Superior Temporal Gyrus	38	6.93	568	[44 22 -30]
Inferior Frontal Gyrus	47	-6.90	560	[-16 20 -20]
Superior Frontal Gyrus	11	-6.83	504	[18 50 -18]
Inferior Frontal Gyrus		-6.83	1072	[34 26 -14]
Medial Frontal Gyrus		-6.82	832	[8 66 12]
Middle Occipital Gyrus		6.80	2000	[-46 -76 -16]

Middle Occipital Gyrus	19	-6.77	472	[-34 -96 10]
Frontal Lobe		6.75	912	[8 44 -24]
Superior Frontal Gyrus		-6.67	280	[-24 52 -18]
Inferior Frontal Gyrus		-6.55	1168	[46 40 6]
Superior Frontal Gyrus	10	6.55	336	[28 62 6]
Parahippocampal Gyrus	36	-6.48	2016	[-36 -32 -24]
Middle Temporal Gyrus		6.47	1096	[52 -74 24]
Superior Frontal Gyrus	8	6.43	2168	[-24 34 56]
Extra-Nuclear		-6.41	288	[0 8 4]
Lateral Ventricle		6.36	344	[-22 -44 8]
Superior Temporal Gyrus	38	-6.32	1032	[-36 16 -22]
Extra-Nuclear		-6.31	2208	[20 -42 20]
Fusiform Gyrus		-6.31	728	[-28 -62 -14]
Cuneus	19	-6.28	280	[-26 -94 24]
Transverse Temporal Gyrus	42	6.28	672	[-66 -18 12]
Paracentral Lobule	5	6.27	560	[-4 -48 62]
Inferior Temporal Gyrus		-6.18	640	[58 -32 -20]
Superior Frontal Gyrus	10	6.10	304	[-34 62 22]
Frontal_Inf_Oper_L		-6.00	1040	[-46 16 20]
Cuneus	17	-5.98	1048	[-4 -84 2]
Superior Temporal Gyrus	39	-5.95	776	[-50 -56 8]
Precuneus	7	-5.94	296	[-2 -60 44]
Precentral Gyrus	6	5.93	232	[36 -24 68]
Middle Frontal Gyrus		-5.89	448	[-32 40 -8]
Inferior Frontal Gyrus	47	5.82	480	[52 40 -6]
Inferior Parietal Lobule	40	5.82	600	[-56 -36 30]
Postcentral Gyrus	3	5.67	488	[44 -24 60]
Middle Frontal Gyrus		5.66	336	[42 56 -6]
Supramarginal Gyrus		-5.66	488	[48 -50 28]
Middle Temporal Gyrus		-5.65	504	[-52 -22 -8]
Superior Parietal Lobule	7	5.65	256	[-36 -74 46]
Corpus Callosum		-5.64	344	[-2 -34 6]
Precentral Gyrus	6	5.62	432	[-64 -8 28]
Extra-Nuclear		-5.59	264	[-20 -24 24]
Superior Frontal Gyrus	10	-5.56	256	[24 50 -4]
Middle Occipital Gyrus		5.50	224	[-40 -86 -2]
Superior Temporal Gyrus		-5.49	824	[60 -48 8]
Frontal Lobe		5.45	296	[24 -16 28]
Precuneus	7	-5.41	488	[18 -56 58]
Posterior Cingulate	29	-5.40	408	[-4 -58 8]
Inferior Frontal Gyrus	45	-5.36	416	[-56 38 4]

Precuneus	7	-5.28	296	[-14 -76 48]
Middle Temporal Gyrus		5.27	264	[48 -60 2]
Anterior Cingulate		-5.26	272	[-2 32 8]
Putamen		5.16	1304	[32 -10 2]
Putamen		5.16	256	[14 10 -8]
Parahippocampal Gyrus		-5.10	256	[24 -46 -10]
Cingulate Gyrus		5.08	472	[-4 -42 30]
Thalamus		5.07	312	[-10 -30 8]
Thalamus		4.96	232	[16 -32 12]
Frontal Lobe		-4.79	280	[28 44 8]

Table A - 2. Decision boundary weight map: ***actual large reward***

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Superior Frontal Gyrus	6	-18.50	264	[-4 -18 78]
Medial Frontal Gyrus		-17.71	264	[0 -16 76]
Caudate		-14.95	1864	[4 4 6]
Parietal_Inf_L		-12.75	1336	[-50 -44 56]
Precentral Gyrus		-12.68	448	[18 -20 78]
Frontal_Sup_Medial_L		-12.64	808	[2 60 30]
Lingual Gyrus		12.46	1104	[-18 -90 -14]
Postcentral Gyrus	1	11.66	296	[30 -34 72]
Lingual Gyrus	17	10.96	2720	[10 -96 -8]
Cuneus	30	10.71	1200	[4 -68 4]
Corpus Callosum		-10.66	336	[0 -42 4]
Postcentral_L		10.54	440	[-30 -36 72]
Paracentral Lobule		10.17	936	[-4 -34 72]
Paracentral Lobule	4	9.79	232	[0 -34 74]
Inferior Frontal Gyrus	47	-9.49	304	[-54 18 -2]
Cuneus	18	-9.38	344	[-6 -100 18]
Cuneus	19	-8.83	672	[4 -88 36]
Extra-Nuclear		-8.82	400	[-2 -26 10]
Middle Frontal Gyrus	10	-8.82	528	[-44 50 22]
Postcentral Gyrus	3	8.56	224	[12 -28 78]
Medial Frontal Gyrus	6	-8.41	552	[2 -4 56]
Lingual Gyrus	18	8.32	224	[-6 -80 -10]
Superior Frontal Gyrus	6	8.23	624	[22 28 60]
Medial Frontal Gyrus	6	-8.22	304	[-2 -6 56]
Superior Temporal Gyrus	22	-7.96	224	[64 -12 4]
Amygdala		6.78	248	[-20 -8 -12]
Superior Temporal Gyrus	39	-6.54	240	[52 -56 10]

Table A - 3. Decision boundary weight map: **anticipation of no loss**

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Medial Frontal Gyrus	6	15.71	5488	[0 -12 76]
Middle Occipital Gyrus		14.08	7040	[42 -74 -16]
Medial Frontal Gyrus	6	13.79	592	[-4 -14 76]
Inferior Parietal Lobule	40	12.48	3872	[44 -42 60]
Cuneus	18	10.92	10160	[-16 -100 4]
Lingual Gyrus	18	-10.90	1896	[-2 -86 -4]
Posterior Cingulate	23	10.26	1088	[-2 -58 12]
Corpus Callosum		10.26	1752	[0 -42 4]
Caudate		10.05	1200	[4 6 4]
Cuneus		9.94	1408	[2 -84 32]
Precuneus		8.92	648	[28 -50 2]
Superior Frontal Gyrus	8	-8.87	1320	[12 44 54]
Cuneus	17	-8.75	1048	[8 -96 0]
Middle Frontal Gyrus	10	8.32	2944	[-46 48 16]
Inferior Frontal Gyrus	47	8.07	232	[-54 20 -2]
Posterior Cingulate	23	7.78	368	[2 -62 12]
Cuneus		7.66	256	[10 -102 18]
Insula		7.56	1144	[44 4 -8]
Cingulate Gyrus	32	7.02	576	[4 12 38]
Parietal_Inf_L		7.00	1720	[-46 -52 56]
Middle Frontal Gyrus	10	6.96	1024	[44 44 22]
Superior Temporal Gyrus		6.82	1568	[54 -58 14]
Medial Frontal Gyrus	6	6.57	296	[2 -4 56]
Middle Frontal Gyrus	8	6.52	392	[-54 14 42]
Postcentral Gyrus		6.30	272	[-6 -44 76]
Precuneus		6.08	656	[-14 -80 38]
Postcentral Gyrus		5.85	336	[-60 -12 18]
Precentral Gyrus		5.80	296	[58 -16 42]
Precuneus		-5.70	216	[40 -76 36]

Table A - 4. Node-specific distribution map: ***neutral facial expressions***

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Inferior Occipital Gyrus		-8.47	1264	[-32 -88 -22]
Inferior Parietal Lobule	39	6.33	298240	[42 -68 40]
Lingual Gyrus	17	-5.42	400	[12 -96 -18]
Superior Frontal Gyrus	10	-5.28	2400	[-28 68 6]
Precuneus	7	-4.99	280	[-2 -60 64]
Superior Frontal Gyrus	10	-4.41	1552	[30 66 2]
Medial Frontal Gyrus	10	3.70	416	[10 64 14]
Caudate Body		3.59	648	[6 4 10]
Fusiform Gyrus	37	3.47	552	[36 -48 -20]
Inferior Temporal Gyrus	20	-3.41	296	[-36 0 -50]
Middle Frontal Gyrus		3.34	272	[26 2 68]
Middle Frontal Gyrus	46	3.34	1048	[-48 30 22]
Inferior Frontal Gyrus		3.28	1144	[48 8 36]
Medial Frontal Gyrus	10	-3.21	288	[14 62 4]
Superior Temporal Gyrus		3.18	440	[-44 10 -22]
Middle Frontal Gyrus		3.12	384	[36 44 18]
Middle Temporal Gyrus		3.03	1168	[-52 -16 -6]
Precentral Gyrus		2.95	312	[48 -6 48]
Inferior Parietal Lobule		2.94	344	[52 -24 30]

Table A - 5. Node-specific distribution map: ***actual large reward***

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Postcentral Gyrus		15.68	21040	[30 -38 70]
Lingual Gyrus	17	14.33	7672	[-12 -94 -8]
Lingual Gyrus	17	13.46	4504	[6 -100 -8]
Paracentral Lobule	4	11.18	7024	[0 -36 74]
Postcentral Gyrus	40	10.14	1344	[64 -24 14]
Corpus Callosum		9.53	3224	[0 -38 4]
Thalamus		8.77	1952	[8 -4 14]
Superior Temporal Gyrus	38	7.75	584	[-52 18 -8]
Superior Temporal Gyrus	38	7.68	512	[52 16 -10]
Precuneus	7	7.55	352	[-2 -62 60]
Postcentral Gyrus	7	7.46	336	[-20 -56 68]
Middle Frontal Gyrus		6.99	888	[-50 6 46]
Posterior Cingulate	30	6.59	1864	[6 -62 4]
Precuneus	7	6.53	240	[-2 -80 44]
Paracentral Lobule	31	6.37	400	[2 -32 48]
Cuneus	18	-6.36	320	[6 -100 18]
Inferior	40	6.28	416	[62 -46 22]
Amygdala		6.25	304	[20 0 -18]
Cingulate Gyrus	32	6.24	496	[-2 18 38]
Sub-Gyral		6.16	520	[36 -68 -16]
Fusiform Gyrus		5.89	272	[-42 -56 -20]
Lingual Gyrus	18	-5.67	232	[2 -80 -2]
Inferior Parietal Lobule	40	5.36	272	[-40 -58 54]

Table A - 6. Node-specific distribution map: ***anticipation of no loss***

	Brodmann Area	Weight	cluster size (mm^3)	peak voxel (MNI)
Lingual Gyrus		-12.30	10920	[-14 -98 -10]
Precentral Gyrus	6	-11.27	34744	[36 -26 68]
Lingual Gyrus		-10.00	9192	[26 -94 -10]
Extra-Nuclear		-8.44	14216	[-6 0 4]
Inferior Frontal Gyrus	47	-8.29	6248	[-52 18 -6]
Superior Parietal Lobule	7	-8.01	9440	[-36 -62 58]
Cuneus	18	-7.53	10728	[-2 -82 22]
Superior Temporal Gyrus	38	-6.74	1984	[50 16 -12]
Superior Frontal Gyrus	6	-6.53	5080	[-4 -8 74]
Precentral Gyrus		-6.03	736	[-30 -26 72]
Superior Temporal Gyrus	22	-6.01	1976	[58 -60 14]
Middle Frontal Gyrus	6	-5.95	1040	[-32 -2 66]
Lingual Gyrus	18	5.88	272	[-2 -84 -4]
Precuneus	19	5.64	352	[42 -74 42]
Superior Occipital Gyrus		-5.18	1680	[-34 -80 24]
Middle Frontal Gyrus		-4.96	1784	[-44 48 12]
Cuneus	19	4.95	744	[16 -100 22]
Superior Frontal Gyrus	6	-4.86	440	[-2 12 62]
Paracentral Lobule	31	-4.76	384	[2 -18 50]
Sub-Gyral		-4.56	648	[42 -8 -14]
Precentral Gyrus		-4.29	256	[56 6 10]
Inferior Frontal Gyrus	9	-4.06	632	[54 4 36]
Cuneus	19	-4.03	224	[28 -84 26]

Acknowledgements

First, I would like to thank my primary supervisor, Prof. A.J. Fallgatter, who gave me the opportunity to conduct my research in his lab and who – with his optimism and confidence in me and my ideas – enabled the exploration of new paths beyond the beaten track. Without his openness to new suggestions combined with his constructive criticism, this work as well as all of my research over the past years would not have been possible. In this context, I would also like to thank my two co-supervisors, Prof. K.-P. Lesch and Prof. M. Heisenberg, who – with their encouragement and their curiosity – motivated me and showed interest and confidence in my work far beyond the ordinary.

Special thanks also go to the Graduate School of Life Sciences (GSLS) for giving me the unique opportunity to develop my own ideas and for providing the organizational structure to successfully pursue them. In particular, I am indebted to Mrs. Lichtlein for providing excellent support in the preparation of my research stay in London and to Dr. Blum-Oehler and the rest of the GSLS staff for rescuing me from every red tape crisis I caused.

Moreover, I would like to thank Prof. M. Brammer and Dr. J. Mourão-Miranda for having me in London and introducing me to the thrilling field of machine learning and the vibrant communities at UCL and KCL. In particular, I thank A. Marquand from KCL for his unparalleled combination of modesty and ingenuity and for more than generously sharing the tricks of the trade with me.

I thank all of my colleagues and friends from the work-group “Psychophysiology and Functional Imaging” in Würzburg who have – with countless comments and suggestions – contributed to every stage of my project. Specifically, I am grateful to Thomas Dresler with whom I established the fMRI measurement protocols at the Research Center Magnetic Resonance Bavaria (MRB) and together with whom I spent countless hours acquiring most of the data on which this work is based. It would have been nowhere near as pleasurable without his talent to imitate virtually every person he has ever seen. Furthermore, I thank Dr. M. Plichta for introducing me to fNIRS and fMRI methodology and to a creative while rigorously principled way of analysing imaging data. I thank Dr. A.-C. Ehlis – beyond anything professional that would surely justify more than one invitation to dinner – for being much more than one could hope for in a friend. Also, I am indebted to the physicians and the staff at the Clinic for Psychiatry, Psychotherapy and Psychosomatics in Würzburg who were a great help in recruiting patients.

Last, but by no means least, my special thanks go to my family – in particular to my parents and my partner Lisa – without whose unconditional support, encouragement, and love this work and much more would not have been possible.

Affidavit

(Eidesstattliche Erklärung)

I hereby declare that my thesis entitled

Integrating neurobiological markers of depression: an fMRI-based pattern classification approach / *Integration neurobiologischer Marker depressiver Erkrankungen mittels fMRT-basierter Musterklassifikation*

is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I verify that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Würzburg, March 9, 2010 ……

Signature

Curriculum Vitae

Personal information

Address	Tim Hahn Tiroler Straße 74 60596 Frankfurt am Main
Email	Hahn_T@klinik.uni-wuerzburg.de
Date of birth	July 19, 1981
Place of birth	Lennestadt, Germany
Nationality	German

Education

1988 – 1992	Christine Koch Grundschule, Bamenohl
1992 – 1998	Gymnasium Maria Königin, Lennestadt
1998 – 1999	Greenfield High School, Greenfield, Wisconsin
1999 – 2001	Gymnasium Maria Königin, Lennestadt

Scientific training

2001 – 2007	Undergraduate and graduate studies of psychology, Philipps-Universität Marburg
since 2007	Fellow of the International Graduate School of Life Sciences, Würzburg

Würzburg, March 9, 2010